Evaluating the User Experience of an Augmented Reality Application Using Gaze Tracking and Retrospective Think-aloud

Tommi Pirttilahti

University of Tampere
Faculty of Communication Sciences
Human-Technology Interaction
M.Sc. thesis
Supervisor: Päivi Majaranta
June 2017

University of Tampere
Faculty of Communication Sciences
Degree Programme in Human-Technology Interaction
Tommi Pirttilahti: Evaluating the User Experience of an Augmented Reality Application Using Gaze Tracking and Retrospective Think-aloud
M.Sc. thesis, 54 pages, 3 index and 10 appendix pages
June 2017

Gaze tracking has previously been used to evaluate usability, but research using gaze tracking to evaluate user experience has not been conducted or is very limited. The objective of the thesis is to examine the possibility of using gaze tracking in user experience evaluation and providing results comparable with other forms of user experience evaluations. A convenience sample of ten participants took part in an experiment to evaluate the user experience of an augmented reality application. Gaze tracking was used as a cue to help participants recall their user experience in a retrospective think-aloud. Participants also filled in a user experience questionnaire and were interviewed about their experience of using the application. The results of the experiment suggest that gaze tracking can be used in measuring user experience when combined with the retrospective think-aloud method. The quotes generated can be used to establish which features or qualities of the application affected the user experience of participants. The method establishes a basis for further research for using gaze tracking to evaluate user experience.

Key words and terms: Gaze tracking, User experience, Retrospective think-aloud

Acknowledgements

First of all, I would like to express my gratitude to Päivi Majaranta for supervising my thesis and offering me great advice throughout the process. She introduced me to the basics of gaze tracking, which was essential for the successful completion of my study. Furthermore, she was always ready to guide me whenever I faced obstacles.

I am grateful to Markku Turunen for his insightful advice and comments given during multiple occasions throughout the development of the thesis and its groundwork.

I would like to thank Deepak Akkil for his help and assistance during multiple occasions including helping with technical difficulties faced while developing the study.

I am also immensely grateful to Yaniv Steinberg for assisting in my user testing. His help was very much appreciated and made a challenging setup possible.

I also acknowledge the contribution of Jari Kangas for a valuable discussion on the interpretation of my results.

Last but not least, I thank the cooperation of Delta Cygni Labs with my thesis. I especially acknowledge the help received from Boris Krassi and Sauli Kiviranta in helping to formulate the structure of the testing for their product.

Contents

1. Introduction

The growing availability and usability of eye tracking technologies has meant that an increasing amount of user research is now done using the help of eye tracking. The likely reason is that eye tracking offers unique possibilities to various fields of science as well as commercial and industrial fields. Eye tracking as its name suggests is the act of tracking the physical position of the eye in order to determine its movement and direction of visual attention. (Romano Bergstrom & Schall, 2014, p. 1-3)

To avoid misconceptions between eye tracking and gaze tracking the following thesis will focus on gaze tracking as the method used, but also acknowledges eye tracking as the wider concept behind gaze tracking. That said eye tracking will be defined as the overall act of tracking a subject's eye movements with the use of technological appliances. Gaze tracking on the other hand will be defined as the act of tracking the direction of the subject's line of sight and the spots associated with the point of focus, involving both movements and fixations of the eyes.

Gaze tracking has been used to analyze gaze behavior since it was first invented. It has more recently also been used as means of control or manipulation in various applications, such as controlling graphical user interfaces (Majaranta and Bulling, 2014). This thesis, however, will focus on the analysis purposes of gaze tracking and how it can be used to benefit the development of products or tools. Gaze tracking as means of analyzing the ease of use of interactive systems has a long history dating back to the mid-20th century, when it was first used to analyze the cockpits of fighter planes (Romano Bergstrom & Schall, 2014, pp. 9).

User testing has developed since it was first established and nowadays the term often used for testing the ease of use of a product is usability testing. Usability can be defined as: "Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use." (International Organization for Standardization, 2010a). The process of testing a product, system, or service is often referred to as usability evaluation or usability testing. These processes can either include gaze tracking or not, but recent developments in both the quality and price of gaze trackers has made it more accessible for researchers to include gaze tracking in user testing. Therefore, gaze tracking has gained popularity in the recent years when conducting usability evaluations. The additional benefit of gaze tracking in usability evaluations depends on multiple factors, including form of data that is being monitored, type of product, system, or service, and whether the right kind of gaze tracker is used.

Usability evaluations, however, do not answer all the questions the researcher might have on the product, system, or service. One of the aspects that usability evaluation does not answer is what the user experience of the product, system, or service is. Therefore,

user experience evaluations have been developed to address these issues. User experience is not easily definable, however, the definition: “a person's perceptions and responses that result from the use and/or anticipated use of a product, system or service” (International Organization for Standardization, 2010b).), has been offered and is one way of looking at user experience, but is not commonly accepted as one.

Contrary to previous literature on gaze tracking and usability, previous literature on gaze tracking and user experience is scarce. Additionally, in many instances it can be argued that the literature is in fact measuring usability disguised as user experience (e.g. Bojko, 2005; Djamasbi, 2014). This might also in part be due to the lack of a commonly accepted definition of user experience. However, to the best of my knowledge there has not been any literature so far that would specifically try to distinguish usability evaluation and user experience evaluation in gaze tracking research.

The benefit of gaze tracking in usability evaluations is widely accepted. Therefore, investigating the benefit of gaze tracking in user experience evaluations can add to the usefulness of gaze tracking and enable new ways of investigating the user experience of users. Thus, understanding the limitations of gaze tracking when analyzing the user experience is vital in developing an accurate method for evaluating user experience.

This thesis focuses on the development of a method to evaluate the user experience of products using gaze tracking as an additional means of gaining user insight. It discusses the challenges involved in using gaze data to accurately interpret the subjective experiences of users and how these challenges were taken into consideration. The challenges involved with measuring user experience lead to the outcome of using gaze as a cue in retrospective think-aloud.

A user experience study was created in collaboration with Delta Cygni Labs to evaluate their remote collaboration application Pointr, which uses augmented reality and a form of video calling to enable users to instruct other users. The first research questions that is answered by the study is:

1. Can gaze tracking be used to aid in the measurement of user experience of digital products?

Following a confirming answer of the first research question, another question is asked of the usefulness of such a method:

2. Are there benefits using methods that combine gaze tracking and user experience in comparison to other forms of user experience measures?

The thesis will continue with two chapters reflecting the background of the research area more in-depth. These chapters will lead towards the method of using gaze tracking

in combination with retrospective think-aloud, which was designed based on previous work. Afterwards, the results of the study will be presented and analyzed.

2. Gaze tracking and usability

The idea of gathering user insight by knowing where the user is looking is fascinating. Now thanks to eye tracking technologies, various methods of collecting data from users' eyes have been developed to complement previous methods of user research, such as think-aloud methods or other forms of contextual inquires. For instance, the usefulness of gaze tracking in usability studies has been well documented (e.g. Holmqvist et al., 2011; Romano Bergstrom & Schall, 2014). Despite the relative infrequency of eye tracking it has a long history as a method of studying human behavior (Holmqvist et al., 2011).

2.1. History of eye tracking

Eye tracking dates back to the late 1800s where the first tools used to measure eye movement were highly intrusive (Holmqvist et al., 2011, p.32). Some of the earliest eye trackers used in 1898 involved inserting a Paris ring, attached to a mechanical lever, into the subject's eye while the participant's eye was anaesthetized with a solution inclusive of three per cent cocaine (Delabarre, 1898 as cited by Holmqvist et al.). To the relive of test participants, Dodge and Cline introduced the method of photographing the reflection of an external light source from the eye's fovea (Dodge and Cline, 1901 as cited by Holmqvist et al.). However, researchers continued to use invasive techniques, some involving apparatuses similar to today's contact lenses (Romano Bergstrom & Schall., 2014). Paul Fitts and his colleagues (1947 as cited by Holmqvist et al.) studied the eye movements of fighter pilots using film based eye tracking. In the 1960s the video based eye tracker became more widely used which led to its development. The downside was that the video based eye trackers remained invasive by requiring the participants to have their heads stuck in one position while biting onto a mouth piece in order to keep their head in one place. In the 1990s the modern eye trackers were first introduced (Romano Bergstrom & Schall, 2014). This meant that eyes could now be tracked without compromising the comfort of the participant and allowed for a more natural interaction (Duchowski, 2007). This led to researchers favoring the non-invasive forms of eye trackers, specifically those based on video, and enabled the use of tracking the eyes even in real time, which further increased the potential applications of eye tracking (Majaranta & Bulling, 2014).

2.2. Anatomy and functionality of the human eye

The human eye is different to most animals in its appearance. Whereas most animals have a dark eye, likely to prevent predators of knowing where they are looking and vice versa, human eyes have a white eye ball, which makes it easier to know the direction of their gaze. The eye is surrounded by six muscles that allow the eye to move with three degrees of freedom. One set of muscles allows the eye to move horizontally, another set

allows the eye to move vertically, and the third set allows for rotational movement. (Drewes, 2010) The human eye is built similar to a camera lens, the outer visible parts consist of a cornea which covers the eye, a sclera, a diaphragm called the iris (see Figure 1), which enables the eye to change aperture, and a lens with a pupil to let light through (Drewes, 2010; Forrester, Dick, McMenamin, Roberts, & Pearlman, 2015).

The outer part of the eye controls the amount of light passed through to the inner part of the eye, the retina. The retina consists of light sensitive rods and cones. The fovea is located at the center of the retina and is the only part of the eye that sees accurately. From the retina, the incoming light is transferred into a picture and sent to the brain through the optic nerve. (Forrester et al., 2015)

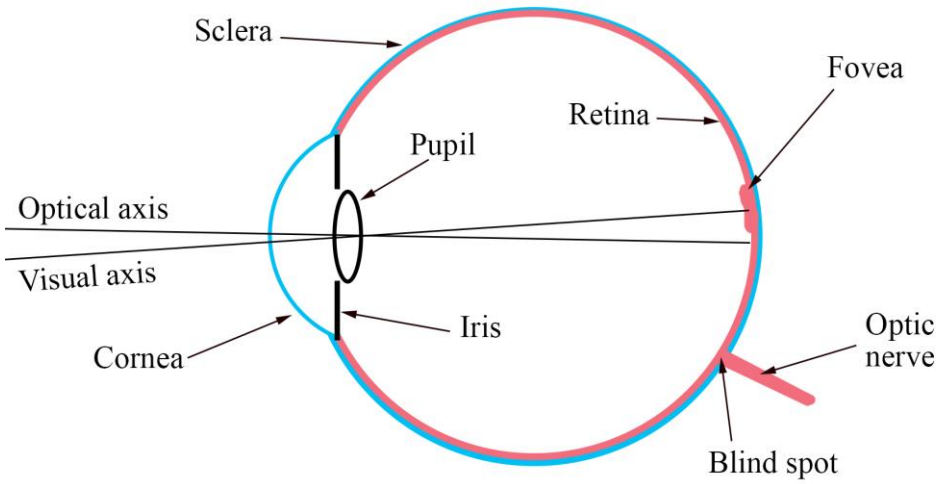

Figure 1. Illustration of the eye. Adapted from Drewes, 2010.

The eye is surrounded by six muscles, which are responsible for the movement of the eye. Of these muscles, two are used for sideways movement, two for up and down movement, and two for "twist" of the eye. To enable humans to see clearly, with only the small point of focus (fovea), the eye moves rapidly to generate a holistic picture of what is seen. (Duchowski, 2007) These rapid movements are called saccades, which are very fast movements, taking 30-80ms to complete and therefore are considered times at which vision is practically blind. Fixations on the other hand are a state at which the eye

remains relatively stable. The combination of the two can be understood as the basic way in which the brain generates the image that we see and are used as the basis of gaze tracking. (Holmqvist et al., 2011)

2.3. Gaze trackers

Gaze trackers are used to estimate the direction of gaze of a person. The traditional gaze trackers can be divided into intrusive and non-intrusive gaze trackers. (Morimoto & Mimica, 2005)

2.3.1. Intrusive gaze trackers

Intrusive eye tracking techniques are usually regarded as more accurate. One of the most traditional eye tracking techniques is inserting a contact lens or a coil into the user's eyes. These approaches are generally very accurate but also extremely intrusive. (Marimoto & Mimica, 2005)

Electrooculography (EOG) is an eye movement measurement approach, that uses electrodes that are placed around the eye to measure small differences in skin potentials (Marimoto & Mimica, 2005). The estimation of gaze with EOG is based on the changing potential of the retina (back) and the cornea (front). The retina has a negative potential and the cornea has a positive potential. When the eye moves to right the potential of the right-side electrode increases and the potential of the left side electrode decreases. The estimation of the new gaze angle is then made in relation θ to the facing direction of the face and the angle the potentials make (see Figure 2). (Manabe, Fukumoto, & Yagi, 2015) With recent technological developments, EOG has also been made non-invasive (e.g. Ishimaru, Kunze, Uema, Kise, Inami, & Tanaka, 2014), but such advances are mainly for research and development purposes and are not commercially available.

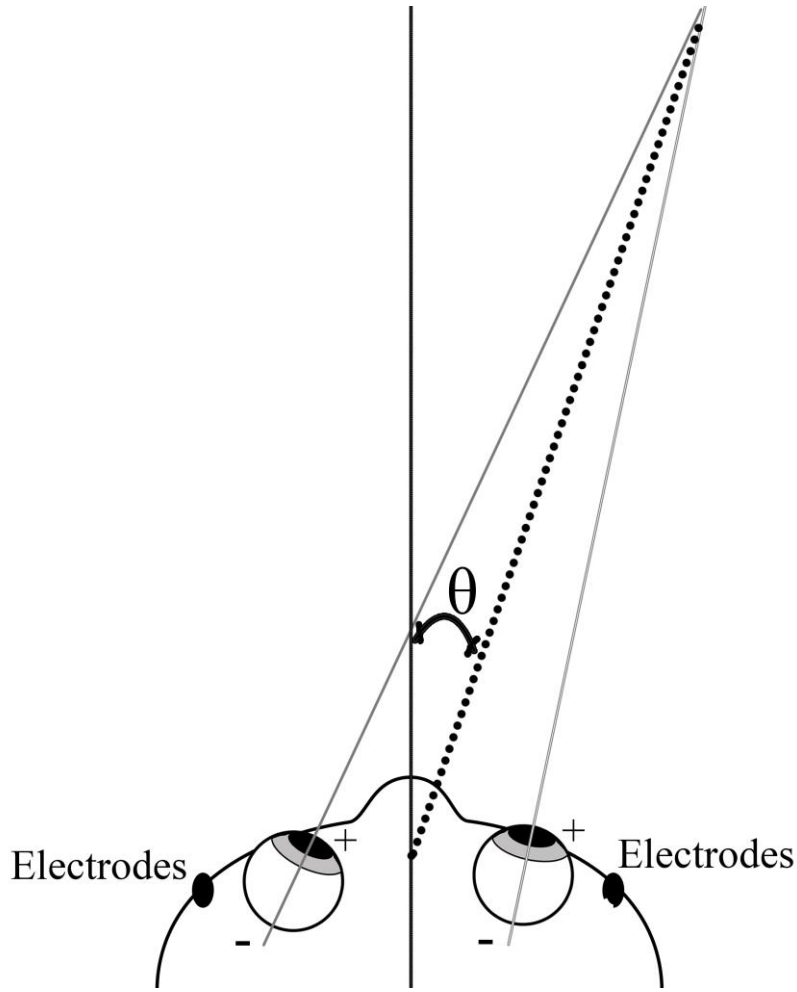

Figure 2. EOG eye angle estimation illustration. Eye rotation changes the potential. Picture is not in scale. Adapted from Manabe, Fukumoto, & Yagi, 2015

2.3.2. Non-intrusive gaze trackers

The commonly used alternative to intrusive gaze trackers are camera based gaze trackers (Morimoto & Mimica, 2005). Camera based gaze trackers are typically cameras that are placed in front of the user. They measure the eye movement of the participant by analyzing the data they receive via images from the camera. There are several different ways of tracking the eyes with camera based gaze trackers, but some of the ways are used more. Holmqvist et al. (2011) use the term pupil-and-corneal-reflection method to describe one of the ways gaze movement is measured using a camera. This technique uses the reflection of the pupil and cornea to determine the direction of the gaze (see Figure 3). Using both the pupil and cornea to determine gaze, the possibility of small movements is preserved for the participant. (Holmqvist et al., 2011)

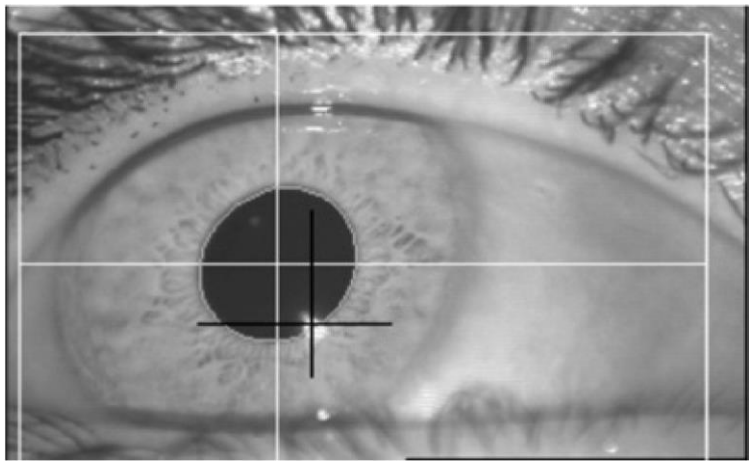

Figure 3. Pupil-and-corneal-reflection system, after properly identifying the pupil. (Retrieved from Homlqvist et al., 2011)

Infrared lights are used in many commercially available gaze trackers to light up the pupil (Holmqvist et al., 2011). This then allows the pupil to be separated from the iris, with better accuracy, and tracked by the camera, without illuminating the user's eyes with light visible to the eye (Morimoto & Mimica, 2005). The typical procedure of tracking gaze when slight head movement is expected can be divided in three steps. The first step is where the camera captures a picture and sends it for analysis. In the next step the picture is analyzed and the center of the pupil is calculated. Finally, the geometrical calculations combined with data from a calibration procedure are used to map the position of the gaze onto the actual stimuli. This is done by comparing the position of the pupil and the corneal reflection and calculating the relative distance between the two at various calibration spots. (Holmqvist et al. 2011) By tracking the reflections from the eyes the system can be made non-intrusive to the participant (e.g. a camera set up in front of the user on the desk). These commercial gaze trackers are usually disguised as black boxes (see Figure 4), probably because of how participants might react to having an easily recognizable camera in front of them, while they perform studies.

Figure 4. Tobii Pro X2 gaze tracker. Tobii AB (2017a)

There are some forms of slightly invasive camera based gaze trackers, such as ones that require the user to stay in one position and therefore use chin rests or other forms of head movement restriction tools. These trackers are generally more accurate, but also restrict much of the natural movements of the participant. Another invasive camera based gaze tracker is the head worn gaze tracker. This gaze tracker is worn on the head to track the gaze of the user in the real world. The tracker enables users to move, more or less, freely in the real world without moving away from the area of gaze tracking, because the camera follows the head movements of the participant (see Figure 5). (Cooke, 2005; Drewes, 2010; Holmqvist et al., 2011; Morimoto & Mimica, 2005)

Figure 5. Head worn gaze tracker. (Tobii AB, 2017b)

2.4. Gaze tracking accuracy and calibration

Gaze trackers are usually compared using accuracy and reliability as measures of data quality (Holmqvist et al., 2011). Accuracy is the property of gaze trackers, which refers to the distance between the actual gaze location and the recorded position (x, y), whereas the precision refers to the reliability of getting accurate readings from fixations of the eye (Nyström, Andersson, Holmqvist, & van de Weijer, 2013).

Accuracy of gaze tracking data depends on various factors, starting from the chosen tracking system, spanning to the proper use of the apparatus. Generally intrusive gaze trackers are technically more accurate (Morimoto & Mimica, 2005), but they are by definition intrusive and therefore they cannot be used for the majority of gaze tracking research. Intrusiveness restricts the natural behavior of participants, causing potential bias which might not show in the data. Therefore, even when the reported accuracy of one search coil was reported as approximately 0.08° (Robinson, 1963 as cited by Morimoto & Mimica, 2005), the results might still be biased, due to the extremely intrusive method.

Like contact lenses/ coils, other forms of intrusive methods are also generally more accurate in comparison to non-intrusive methods. However, the unmeasurable bias that results from even EOG type measures, with only sensors attached to the sides of eyes, might be extreme for participants that are not used to sensors that are attached to their bodies.

Non-intrusive gaze trackers vary largely in their accuracy, but the offset (Holmqvist et al., 2011) of the accuracy can be calculated and taken into account. The distraction bias involved, which reflects the overall performance can be considered minimal in comparison.

Video-based gaze trackers must be calibrated in order to measure gaze accurately. This is done by setting the offset and precision of each participant at optimal levels using various points of reference. This can be done using either manual calibration procedures, where each point is calibrated individually by the moderator of the study or automatically, where the computer automatically measures various spots from the screen to calculate the accuracy quickly and with ease. (Nyström et al., 2013)

The way calibration is usually done in practice by having the participant look at certain parts of the screen and then having the computer calculate the correct reading from the angle the participant is looking at, from several different places. For most commercial gaze trackers, this is accomplished by having the computer screen present points on the screen, where the participant needs to look at. (Goldberg & Wichansky, 2003)

Even after calibration, issues with accuracy might occur due to physical properties of the participant’s eyes, such as small pupil size or eye lids that cover part of the pupil (Goldberg & Wichansky, 2003). Other possible physical disturbances for the eye

tracking system are eye glasses, which might cause incorrect reflections, that the gaze tracker then reads and causes inaccuracy in the data. Therefore, analyzing gaze tracking data needs special caution, especially in cases deviating from the average. However, it is also important not to separate deviating cases just because they deviate, because such data can potentially contain important information, which was not apparent from the other data.

2.5. Analysis with gaze tracking

Type of gaze tracker is chosen based on the need for analysis. Most modern gaze analysis methods only consider using non-intrusive analysis methods (Chennamma & Yuan, 2013; Cooke, 2005). When considering analyzing something with gaze tracking there are important factors to consider. First of all, is the method effective and will the method provide new insight into the research question. Next, what are the possible biases involved and how can they best be avoided. Consequently, when the gaze itself is not the main research objective it is arguable to avoid using intrusive methods, this is also the case with the study involved with this thesis.

The eye-mind hypothesis (Just & Carpenter, 1976), where attention was fixated on what the eye was looking at, was a dominant view among researchers for a long time. This idea, however, was questioned by Posner, Snyder, and Davidson (1980) introducing the concept of attentional spotlight, where vision moves around and only registers important objects or other important features in the line of sight. Attention can therefore not be accurately interpreted from gaze alone, but gaze does indicate the direction of attention and therefore acts as an important cue. To build on this idea there are theories on attention that accompany the results such as the feature integration theory (Treismann & Gelade, 1980), which states that first the overall shape of objects are analyzed and then features are added to the mental picture if attention is focused on the object. On the other hand, it is not possible to attend to one thing and look at another thing (Hoffman & Subramaniam, 1995). When considering the implication that attention sets on interpreting the meaning of gaze, it becomes further arguable that minimal disturbance should be placed on the participants in gaze analysis testing that is interested in attention.

Furthermore, analysis of the gaze data also needs careful consideration. The accuracy of data is subject to many layers of analysis, starting from raw data and moving to computing specific metrics (Goldberg & Wichansky, 2003). After the data has been collected and put into understandable form, it still needs to be categorized somehow, which again involves multiple steps. The first step is to decide what is important. Should all the data available be used, or how should the used data be collected. Next the important data needs to be categorized either into qualitative or quantitative categories, which will then determine the type of analysis that will be done. This also depends on what is seen as important and there are no clear answers, but

qualitative data can for example be categorized based on different quotes, behaviors, or actions and quantitative data can for example be categorized into time taken, number of errors, or number of phases before completion.

2.6. Gaze tracking in usability evaluation

Gaze tracking has been used to evaluate usability for decades. The earliest forms of usability testing with gaze tracking can be argued to have happened in the 1940's, when Paul Fitts studied fighter pilots' eye movements in order to improve the cockpits of the airplanes (Romano Bergstrom & Schall, 2014).

Usability as a term is not easily explained. The ISO definition is "Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use." (International Organization for Standardization, 2010a). However, Abran, Khelifi, Suryn, & Seffah (2003) argue that there are numerous definitions for usability and different fields use different ones. Usability defined by Nielsen (2012) is a quality attribute of how easy something is to use. It is defined by five quality components:

- Learnability - How easily users are able to use the system during their first encounter with it.
- Efficiency – After learning, how efficient is accomplishing tasks.
- Memorability – How well can the users use the tasks they accomplished before, while using the system during another time.
- Errors – Do users make a lot of errors, while using the system.
- Satisfaction – Are users satisfied with using the system (Nielsen, 2012).

Usability evaluations can be designed to measure the usability of the whole product or features of the product. The objective of usability evaluations, however, is always to evaluate the need for changes in the product to better fulfill the requirements for users' ease of use of the product. (Chowdhury & Chowdhury, 2011)

The scope of usability studies extends from physical to digital products and environments. In the past usability studies were always about inferring the ease of use of physical products or environments, such as the physical space where a driver in a car interacts with the car, including the driver's seat, steering wheel, and dashboard. By analysing the usability of such physical spaces and the products included, such as the steering wheel, the cars usability could be improved. Nowadays more and more usability evaluations are conducted on digital products and environments and the focus in many are primarily on the user interface of the product. The user interface is the part of the product that the user interacts with. Within digital products the user interface is most commonly graphical, but more and more embedded interfaces are emerging in ubiquitous computing (Dourish and Bell, 2011). Graphical interfaces being the most commonly used are interfaces, where the user needs to interact with a graphical representation of a control panel of sorts, the controls, however, are operated either by

touch or some form of separate controller such as the mouse. When conducting usability studies, it is determined if the chosen way of interaction is a usable way of interacting with the product or environment.

The general way usability studies are conducted is by developing a research question and operationalizing it into measurable tasks. The way these tasks are analyzed is then decided and proper participants are recruited based on the products' potential/existing user base. The evaluation is then conducted and data is collected based on the preplanned method. After conducting the evaluation, the data is analyzed and recommendations for changes in the product are made if relevant. (Chowdhury & Chowdhury, 2011)

Data collection or method of giving tasks can vary largely in usability evaluations. Data can be collected by having the user fill in questionnaires or researchers can interpret the meaning of what the participants say or do within a certain time frame or while interacting with a certain feature. In the aforementioned situations, it is hard to determine if the reason the participant acted the way they did or answered the way they did is, because they were distracted or if they notice essential information for tasks (Pretorius, Calitz, & van Greunen, 2005). Given the challenges in knowing what the user is attending to and what they are not noticing at all, gaze tracking can potentially bring new information and more accurate interpretations of the usability issues in products (Ehmke & Wilson, 2007; Pretorius et al., 2005).

Ehmke & Wilson (2007) listed the most common ways of analyzing gaze data in relation to usability evaluation, which are fixation-related, saccade-related, scanpath-related and gaze-related analyzes. With fixation-related analysis the attention is drawn to where the fixations take place and how long they remain fixated on the target. Fixations can tell of where the participant's attention was while accomplishing tasks. When analyzing saccades, attention is focused on how many saccades there are and what the amplitude of the saccades are. Generally, the more saccades there are the more searching is being done by the participant. Scanpath analysis considers the length and direction of scanpaths (i.e. the path generated, usually by software, visualizing saccades and fixations). They are used to analyze the efficiency of search or effectiveness of layout. Gaze-related analysis takes into account how gaze acts overall, whether it dwells or revisits certain areas. This information can then be used to analyze if something causes more confusion. (Ehmke & Wilson, 2007)

2.7. Shortcomings of usability evaluation

The chosen usability evaluation method can affect the outcome of the results. Choosing the right method is therefore important for good data. The problem is that there is no clear way to tell what the right method is. Choosing the wrong method might mean that important functions were not taken into account, or that something important was overlooked, because the method does not give accurate enough results.

Using gaze tracking can help to decrease the amount of important factors that were missed, but gaze tracking is not the right choice for all kinds of usability evaluation. Furthermore, if gaze tracking is used there are still a number of different approaches to what is the correct way of analyzing the gaze data. Choosing the wrong analysis method might mean that the results are biased, or that not enough information is available after evaluating.

Usability evaluations can be essential for good product development; however, usability evaluations alone are not sufficient. The purpose of usability evaluations is generally to find flaws or ways of making the product easier to use. This does not take into account most of what is happening while users interact with a product. Ease of use alone does not mean that the product is good. Take for example, an old mobile phone without a touchscreen or other luxuries provided by modern mobile phones. They are generally regarded as easy to use. This ease of use comes at a cost. The mobile phones in question are good at making calls to another phone, the interface is easy to use, the person using it can easily figure out how to make a call and remembers how to make the call next time also. It is so efficient it only takes a few clicks to make a call. There are close to no errors because of the clunky keypads, and the user is satisfied with how everything works. So why are people not using these old phones anymore? There might be several reasons, however, when looking at the usability evaluation criteria, it seems like the old mobile phone is good and nothing should be changed.

When evaluating a product's usability, the researchers are only evaluating if the product works like they believe it should. What they might be missing is that users would not like to use the product despite it being easy to use. Sometimes people might even prefer some challenge, for example Norman (2013) gives the example of people liking to read normal books or magazines, despite digital versions (ebooks, webpages) being without the hassle of turning physical pages and having book shelves full of books, etc. Furthermore, according to Goldberg and Wichansky (2003), the cognition of the participants is not easy to capture with usability testing. Therefore, usability alone does not give the whole picture.

3. User experience and gaze tracking

User experience has gained popularity among academic researchers and the industry in the human computer interaction community during the recent years (Mirnig, Meschtscherjakov, Wurhofer, Meneweger, & Tscheligi, 2015). This has resulted in user experience being often mixed up with usability (Rusu, Rusu, Roncagliolo, Apablaza, & Rusu, 2015), but these two terms are not equivalent. Usability can be seen as part of user experience and user experience can be seen as complementary to usability, but mixing the terms only results in confusion. The mix up of these terms can in part be due to the missing of a commonly accepted definition (Bevan, 2009).

There are multiple methods of evaluating user experience and new methods are being invented continuously. Exploration of past methods can give a good impression of the possibilities involved in user experience and by using these past methods and linking them with new ideas, new research areas can be created. This chapter will cover how connecting past research on user experience and gaze tracking can be utilized in development of new user experience approaches.

3.1. User experience vs. usability

As stated before, usability is defined as “Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.” (International Organization for Standardization, 2010a). whereas, user experience is defined as “a person's perceptions and responses that result from the use and/or anticipated use of a product, system or service” (International Organization for Standardization, 2010b). The definition used by ISO 9241-210:2010 for user experience while being simple, can be interpreted in many ways.

The importance of separating the two terms is clear when considering the limitations of usability. Usability measures the ease of use of a product, system, or service, but does not consider the user’s experience, other than that of their satisfaction. Satisfaction can be seen as part of user experience; however, it is not a synonym for experience. Experience is a much more abstract term and cannot be measured on a scale of one to seven, like satisfaction can potentially be. User experience can also potentially be high even when usability is poor, this can be seen in poorly designed games, where playing is fun even though the execution of the game is not the best it could be. Hassenzahl (2008) defines user experience by “a momentary, primarily evaluative feeling (good-bad) while interacting with a product or service”, this definition encompasses that even when the user might feel bad at times while interacting with a product, their overall feeling might still be positive. User experience is a rather subjective term, whereas usability can be considered more objective (Hassenzahl, 2008).

Separating usability and user experience does not mean that they cannot complement each other. Moczarny, De Villiers, & Van Biljon (2012) present three different perspectives in which user experience and usability are seen to be linked, the first being that user experience subsumes usability and is therefore the higher category for usability. Another view takes the opposite perspective, where usability is seen as the higher category and user experience is part of the satisfaction component. The third perspective is that the two are separate concepts, but they intersect with common attributes, but also have their own individual attributes (see Figure 6). (Moczarny et al., 2012)

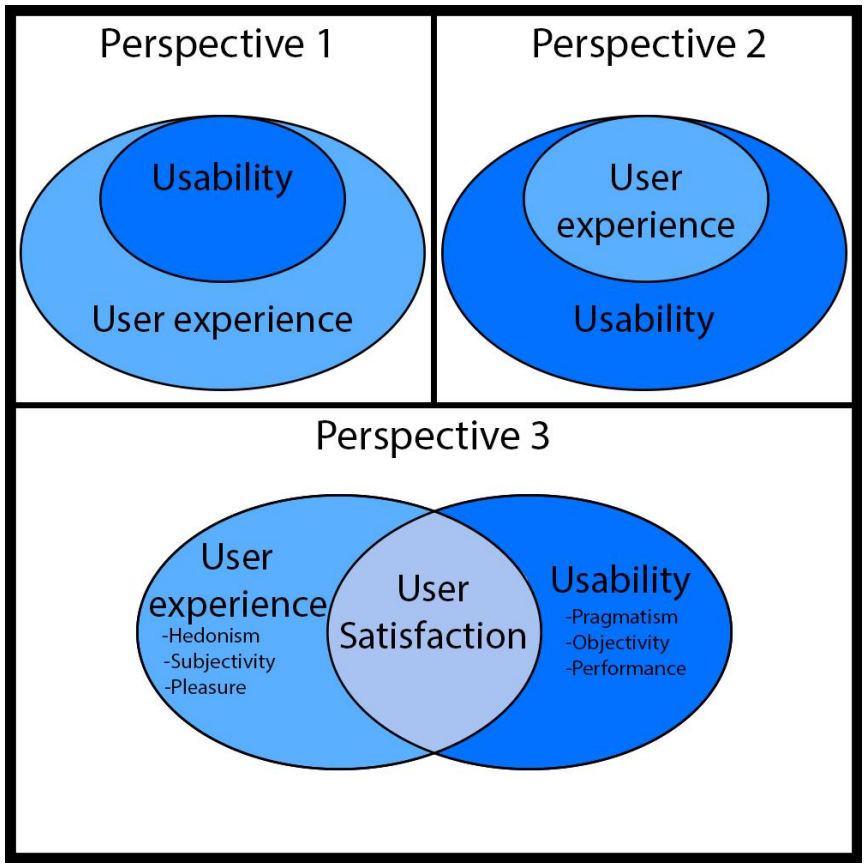

Figure 6. Three different perspectives of the relation between user experience and usability. Adapted from Moczarny et al., (2012).

Hassenzahl (2007), differentiates between dimensions of user experience by using the pragmatic vs. hedonistic model of user experience, where the pragmatic dimension refers to the product's perceived ability to support the achievement of "do-goals", such as "establishing connection" or "finding the right button". The hedonistic dimension refers to the product's perceived ability to support the achievement of "be-goals", such as "being inspired" or "being surprised". The model gives a theoretical way measuring user experience and including elements, which could be argued to be usability, but

should be regarded as part of user experience. By using the model to gain user insight, the need to separate between usability and user experience can be seen as non-essential, nevertheless, the benefits of both are included. The model, however, assumes that pragmatic and hedonistic aspects of user experience are separate and the same product features can have both pragmatic and hedonistic perceptions at the same time. (Hassenzahl, 2007)

3.2. User experience indicators

Due to the abstract nature of user experience it is hard to separate the factors that influence user experience, however, several indicators have been distinguished. Two factors stand out from former literature: affect and aesthetics (Bargas-Avila & Hornbæk, 2011).

In human-technology interaction related literature emotion is usually regarded as both the physical and non-physical experience of feelings. Some prefer to use the term affect (e.g Bargas-Avila & Hornbæk, 2011), that is more commonly used in psychological literature to separate different levels of emotional experience from the overall experience of the commonly used word, emotion (Russell, 1978). Affect is therefore often measured on a subjective level. It includes both the internal state of the person and the consequence of these states (Russel, 1978). By measuring affect, it is possible to interpret whether the experience was positive or not. Positive experience is expected to lead to good user experience, at least if the overall experience was good, and the positive affective states were related to the system, product, or service. It is important to notice that affect is a changing state, and because of its changing nature, measuring it is problematic. Therefore, the size of the gap between measurements of affect, can bias the results (Bruun & Ahm, 2015).

Aesthetics on the other hand, influences the user's experience through thought and behavior (Tractinsky, 2004). Aesthetic information is often what creates the user's first impression (Djamasbi, Siegel, Skorinko, & Tullis, 2011), and reactions to aesthetic stimuli are considered fast (Tractinsky, 2004) meaning that the user might form an impression of a product, system, or service, just by looking at it. The original impression can change, but Djamasbi, et al. (2011) argue that the expectations of users are also changing and users might decide not to proceed with their interaction, after a brief while if the first impression does not satisfy them. Research on aesthetics in human computer interaction has been on the rise, and some researchers have even claimed that what is beautiful is also useable. However, Tuch, Roth, HornbæK, Opwis, and Bargas-Avila (2012) argue that the causation of the claims have been turned around and the case is more likely, what is usable is beautiful. This in part suggests that Hassenzahl's practical and hedonistic theory of user experience (2007) might be a good way of combining different aspects of user experience.

Furthermore, attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty are used in the UEQ (user experience questionnaire) to measure the overall user experience of using a product (Laugwitz, Held, & Schrepp, 2008). Attractiveness can be seen as a measure of aesthetics, whereas the other measures are unique. Perspicuity, dependability, and efficiency can be seen as parts of practical user experience (Hassenzahl, 2007), which measure how easy it is to get acquainted to the product, whether the user feels in control of the interaction, and how much effort the participant must put into the product, respectively. Stimulation, and novelty can be seen as parts of hedonistic user experience (Hassenzahl, 2007), where stimulation is how exiting or motivating the product is, and novelty is how innovative the product appears. (Laugwitz, et al., 2008)

Other indicators of user experience include, fun, immersion, and flow (Harrison, 2008). These remind of the abstract nature of user experience, where numerous factors have an effect, but separating them can be problematic. Furthermore, in certain cases positive user experiences can arise from the perceived usefulness or functionality of the product, as in some augmented reality services (Olsson, Lagerstam, Kärkkäinen, & Väänänen-Vainio-Mattila, 2013). When evaluating user experience, it often might not be enough to use one measure, due to the complexity associated with the term and therefore mixed-method designs are common in user experience evaluation (Law, 2011). These methods usually combine quantitative and qualitative measures, such as task based evaluations and questionnaires. The chosen methods depend on different aspects discussed in the following subsections.

3.3. User selection in evaluating user experience

Given that user experience is all about the user, it is important to consider the impact of user selection. The quality of the results is directly related to the choice of participants. When users evaluate their experience of using a product, service, or system, their background matters. If participants are familiar with the subject of the test, their experience will be different in comparison to someone who is new to the subject (Keskinen, 2015). It is also worth noting that people like different things, therefore it is important to consider if user population and the participants match. Contrary to the ideal situation, variance in subjective liking will always exist.

Additionally, the number of participants that are needed for user experience evaluations vary depending on the wanted analysis method, validity, and replicability (Ritter, 2013). It is important to consider what the type of analysis is going to be before deciding how many participants should be recruited. In qualitative analysis methods, the number of participants can be relatively low in comparison to quantitative analysis methods. With quantitative analysis as the choice of analysis, the number of participants needed will depend on the expected population parameters. These parameters are

usually unknown and therefore the number of participants can usually be based on the requirements of specific statistical tests.

Often when evaluating a product, the chosen method is qualitative. Additionally, the number of participants is often lower for qualitative studies in comparison to quantitative studies. The reasoning for this has generally been established as resource efficiency and that the biggest problems of the product usually arise during the first few evaluation sessions. When evaluating the user experience of a product it usually does not make sense to spend too much time on evaluating the user experience, because the life cycle of any product is usually limited. Therefore, fast and efficient forms of analysis are better suited. This is also considered in the development of the method described in the thesis.

3.4. User experience evaluation methods

Operationalizing different measures of user experience is challenging (Law, 2011). There is no consensus on which methods are to be used and when. One of the main arguments in operationalizing user experience is the reductionism vs. holism debate, meaning is it justifiable to reduce user experience into quantifiable measures or should it always be measured with qualitative holistic measures. On the other hand, taking a strict stance on either side is not beneficial and due to the fact, recent user experience studies have moved their focus from strictly quantitative studies to qualitative and mixed-method studies. (Law, 2011)

User experience is easily interpreted as something static, which does not change, however, this is not the case. A person's initial experience of a product is not necessarily what his or her experience is after interacting with the product for a while, and can still change after time passes. It is also important to notice that the experience might change without further contact with the product. The dynamic nature of user experience needs to be considered when designing user experience evaluations. (Vermeeren, Law, Roto, Obrist, Hoonhout, & Väänänen-Vainio-Mattila, 2010)

Due to the dynamic nature of user experience, it is important to consider where the setup is located. Therefore, the study type needs to be carefully chosen depending on the research situation and wanted outcome. Based on the meta-analysis by Rajeshkumar, Omar, and Mahmud (2013) the different study types used in user experience evaluation were:

- Laboratory/ controlled study – takes places in a controlled environment. Despite its name, a laboratory study might not take place in a space specifically designed for laboratory studies. The strength of laboratory studies is that independent variables can be manipulated and their effect on the dependent variable can be extracted. Their weakness, however, is that controlling too much, might have an effect on the results, due to the unnatural environment.

- Field studies – address the weakness of laboratory studies. They are situated in “the real world”, where the effect of independent variables on dependent variables cannot be accurately controlled.
- Surveys – are used to gather data from users for analysis, by questioning them.
- Expert evaluations – are the most researcher subjective forms of user evaluation and are based on the researcher’s own educated interpretations of the user’s experience. The researcher uses a predefined domain matrix and parameters to determine the degree of different user experience related factors. (Rajeshkumar et al., 2013)

The time frame of the user experience evaluation is also important to consider (Keskinen, 2015). Both the length of the evaluation and the length of the measurement of the experience, should be taken into account (Rajeshkumar et al., 2013). Most user experience evaluations are not longitudinal, because it would take too many resources, thus most user experience evaluation methods are not tailored for longitudinal studies (Keskinen, 2015). Therefore, the need to consider the length of evaluation is mostly limited to considering how long each test lasts, and how that might affect the measurements. The time frame of the measurement needs to be considered differently, if the measure of experience is taken from the overall experience of using the product, system, or service, or if the measure is taken from a specific moment in the testing (Rajeshkumar et al., 2013).

Furthermore, Roto, Law, Vermeeren, & Hoonhout (2011) discuss how the developmental phase or the iterative process of the product, system, or service should also be considered when deciding on a method. If the product is at an early prototype level, it might be a waste of resources to evaluate the user experience of the prototype, with methods that take an hour for each participant or similar. On the other hand, if the product is already used by many users and it still has issues, which cannot be identified by fast resource efficient methods, gathering detailed information might be most beneficial.

Considering all the different options, choosing a method can be very complex. Even after analyzing all the previously mentioned aspects, choosing a specific method can still be challenging. Previously user experience evaluations have been done with numerous different methods and there is no clear way to choose one, which would best fit any specific case (Keskinen, 2015).

Despite the numerous user experience methods available, methods that are typically chosen are those which are not heavy on resources. These include contextual inquiries, interviews, and think-aloud methods. Semi-structured interviews have been used often in user experience research (see e.g. Law, 2011; Rajeshkumar et al., 2013) and the results of this study suggest that it is an effective method of discovering users’

experiences. The think-aloud method is also a widely-used method, both in user experience evaluation as well as usability evaluation. In the method, the researcher asks the participant to think aloud during the experiment, in order to understand what the participant is thinking, while interacting with the product. The reasoning is that by having the participant talk about why they are doing what they are doing, insight into the user's mind is achieved. (Roto, Obrist, & Väänänen-Vainio-Mattila, 2009) For a list of user experience methods see All about UX (2017).

3.5. Think-aloud methods

Thinking aloud is a procedure where participants are asked to verbalize their thoughts by thinking aloud. The objective is to understand what the user is or was thinking during performing tasks. Think-aloud methods are usually divided into two separate methods, the concurrent and the retrospective think-aloud method (Elling, Lentz, & de Jong, 2011).

3.5.1. Concurrent think-aloud method

The traditional concurrent think-aloud method has been widely used in usability testing (Guan, Lee, Cuddihy, & Ramey, 2006). With the concurrent think-aloud, the researcher asks the participant to think aloud during the usability evaluation session (Hyrskykari, Ovaska, Majaranta, Räihä, & Lehtinen, 2008). By having the users vocalize their thoughts during the experiment, the researcher gets insight into the user's mind during the test (Roto et al., 2009). The procedure has been criticized for interfering with the normal thought process of the participants, for example by demanding them to verbalize thoughts that happen faster than they can speak, or interfering with the completion of the task itself due to cognitive load (Nielsen, Clemmensen, & Yssing, 2002). However, using the method can be acceptable depending on the extent of interference and is argued to have minimal effect in certain situations (Ericsson & Simon, 1993).

Ericsson and Simon (1993) present the idea that certain stimuli are easier to verbalize than others, depending on the stimuli's coding in the short-term memory. They present the idea of "levels of verbalization" (p. 79), which has three different levels. The first level being verbalization of thoughts that are thought of as words that can be directly said. The second level are thoughts that need to be interpreted first, to make them verbalizable. In the third level, the person is required to interpret how his or her thought process was accomplished additionally to verbalizing the thought, Ericsson and Simon argue that this is not directly coded into the short-term memory and needs active processing. Even though verbalization might sometimes require additional processing, sometimes thoughts are verbalized to enhance performance, for example in noisy situations (Ericsson & Simon, 1993). Ericsson and Simon (1993) do not suggest that introspection would be an undisputable pathway to all the thoughts of the user, their analysis demonstrates the potential of the think-aloud method.

3.5.2. Retrospective think-aloud method

In contrast to the concurrent think-aloud method the retrospective think-aloud method asks the participant to describe what they did, after the experiment (Eger, Ball, Stevens, & Dodd, 2007). In practice the most often used form of retrospective think-aloud is the stimulated form, where participants get visual reminders of the tasks (Guan et al., 2006). This in part helps with the criticism that the retrospective think-aloud method relies on the memory of what happened or what the participant was thinking, however, it does not eliminate it (Eger et al., 2007). Nowadays, when evaluating user interfaces a recording of the computer screen is often taken and shown to the participant after the tasks have been completed to stimulate the retrospective think-aloud (Elling et al., 2011).

Formerly, the retrospective think-aloud method has been used widely for usability testing (e.g. Bowers & Snyder, 1990; Van Den Haak., De Jong, & Jan Schellens, 2003). The difference between the results of concurrent and retrospective think-aloud methods have been under investigation and the results are usually similar. Both of the methods produce quantitatively the same amount of verbal responses (Bowers & Snyder, 1990). However, in comparison to concurrent think-aloud the retrospective think-aloud is able to elaborate more on the actual thoughts and experiences of the participant, whereas the concurrent think-aloud seems to be better for strictly error based or practical problem involving information, within the user interface (Van Den Haak et al., 2003).

Furthermore, the retrospective think-aloud procedure has been found to be more effective for producing more words about emotional experiences (Petrie, & Precious, 2010). Petrie and Precious (2010) reason that retrospective think-aloud might distract the participants less when thinking-aloud and therefore lets the participant think about their emotional experience and produces more emotional words. Similarly, as for other experiences (Van Den Haak et al., 2003), emotional experiences can be expected to be less about the errors of the user interface and more about the actual emotions that the interface elicits when evaluated with retrospective think-aloud in comparison to concurrent think-aloud.

Considering the meaningfulness of error based emotional responses and other emotional responses, it can be assumed that error based emotional responses are negative most of the time and errors are also negative. Therefore, those emotional responses have limited additional benefit to the evaluation. Whereas, other emotional responses are more interesting when evaluating the user experience of a product, system, or service.

Recently, similar to other forms of usability testing, retrospective think-aloud methods have started including the use of gaze tracking to add cues for stimulated retrospective recall (e.g. Elbabour, Alhadreti, & Mayhew, 2017; Elling et al, 2011; Hyrskykari et al, 2008; Guan et al., 2006). There are however, concerns in using gaze

tracking combined with retrospective think-aloud. These include the observations that the gaze pattern overlaid on the screen can distract the participant (Elling et al., 2011).

Regardless, most studies comparing retrospective think-aloud with and without gaze tracking have concluded that the gaze pattern overlaid is beneficial and can help participants to find issues that might not be noticed otherwise (e.g. Elbabour et al., 2017; Hyrskykari, et al. 2008). Elbabour et al. (2017) concluded that the gaze cued retrospective think-aloud procedure detects more usability problems, but also can take longer depending on the instructions given (e.g. If the participant is allowed to stop the recording or if the recording is slowed down).

By including the gaze path created by software, the user is able to better recall what they were thinking about at any given moment. This should work better, especially for static graphical user interfaces, where the screen might not move and therefore the recording might remain still for a long time while the participant is searching the interface for the next action (Elling et al., 2011).

3.6. Gaze tracking for user experience evaluation

Gaze tracking cannot be directly used for user experience evaluation. Due to the abstract nature of user experience, looking at the gaze pattern of users will not generate information that could be interpreted as specific kinds of user experience. The only user experience that could be argued to be visible is in some cases erratic search of something simple, which could be interpreted as negative user experience, but would still be missing information that the user could elaborate on. Some literature on user experience evaluation and gaze tracking exists, but the literature often mixes user experience and usability and does not make a distinction between them (e.g. Bojko, 2005; Djamasbi, 2014), instead just causes further confusion.

Considering the available research methods in user experience evaluation and combining them with gaze tracking, the think aloud method is, in my opinion, the most potential. The retrospective think-aloud method enables the use of gaze tracking in user experience evaluation without interrupting the testing. Similarly, to usability evaluation combined with gaze tracking and retrospective think-aloud method, the recording is played back to the participant after the test is complete, with scan paths overlaid. This should produce qualitative data on the user experience of the product, service, or system if the correct instructions are given to the participant and the participant feels that they are not judged, but enabled to present their opinion. Therefore, the procedure should be carefully presented to the participant and practiced before the actual test. By specifying to the user what to focus on while expressing their thoughts, the retrospective think-aloud can be argued to produce data that represents user experience. For example, by instructing the participant to elaborate on their experience instead of what they did. Elling et al. (2011) explained to the participants that instead of telling that they clicked link X, to tell why they thought that link X would produce the response they were

hoping for. By also specifying that the interest lies in the participants' liking of the features or more generally of the product, service, or system, it might be more natural for the participant to relate to the experience aspect of user experience.

The area of user experience evaluation with gaze tracking is new and literature is hard to come by. Therefore, adapting the methodology of retrospective think-aloud from usability evaluations presents the best opportunity of altering prior methods to enable the evaluation of user experience. Hence, a completely new method is not required to test the suitability of gaze tracking for user experience evaluation.

As has been established, user experience and gaze tracking have not been combined previously indicating a research gap, which should be taken into account and will be considered in the thesis. Using gaze tracking to evaluate user experience is not without its problems, as mentioned before, direct measurements are not possible. Therefore, mixed approaches are needed and distinctions between usability and user experience need to be made more clear, which I hope to set a stage for with the thesis. Talk of usability methods will not be present, but practical user experience instead. User experience will be measured through appropriate indicators. Acknowledging the separation of these two concepts and paying attention to accurate terminology will be a separating element in comparison to past research.

To summarize, currently usability and user experience are often mixed up and the literature is confusing. By analyzing the nature of user experience, it is possible to differentiate user experience from usability. The process of differentiating these two has started and there are multiple concepts that describe different aspects of user experience. However, when it comes to gaze tracking, research has continuously been about usability (e.g. Ehmke & Wilson, 2007; Poole & Ball, 2004; Pretorius, Calitz, & van Greunen, 2005), regardless of what the purpose has been. There exists a research gap in finding the most suitable method of user experience evaluation using gaze tracking. To pave the way for further research, the thesis will use a mixed method design including gaze tracking and the retrospective think-aloud method.

4. Method

The following chapter describes the participants, apparatus and materials used, experimental procedure of evaluating the software, and statistical analyses used. The purpose of the study was to evaluate the user experience of the user interface of Delta Cygni Lab's Pointr application (Pointr, 2017), using gaze tracking as a new tool.

4.1. Participants

A convenience sample of 10 participants was collected for the study (see Table 1 for the demographics). English was used as the language of instruction for all the participants attending, despite their mother tongue. All, except one participant was between 20-29 years old. Three of the participants used eye glasses or reported other complications involving their vision. Three of the participants had one previous experience with eye tracking, none had multiple experiences with eye tracking. Two of the participants had used skype or skype like applications as remote collaboration tools. It is unclear how participants understood the question (see Appendix D for background questionnaire). None of the participants had used the application before.

Age group	Gender	Vision	Familiarity with eye tracking	Used other remote collaboration tools	Used Pointr before
20-29	Female	Complication/ Corrected	Yes	No	No
20-29	Female	Normal	No	No	No
30-39	Male	Normal	No	Yes	No
20-29	Female	Normal	No	Yes	No
20-29	Female	Normal	No	No	No
20-29	Female	Normal	No	No	No
20-29	Female	Normal	Yes	No	No
20-29	Female	Complication/ Corrected	No	No	No
20-29	Male	Normal	Yes	No	No
20-29	Male	Complication/ Corrected	No	No	No

Table 1. Participant demographics

4.2. Apparatus and materials

The materials used in the study were the following in their order of appearance. First of all, an informed consent form (see Appendix A) was used to get the participants' consent to participate. For the experimental setting, a mouse with a USB-cord, a web camera with a USB-cord, a micro-SD memory card, a HDMI cable, and a micro-USB charger were used with a Raspberry Pi 3 Model B minicomputer (see Figure 7).

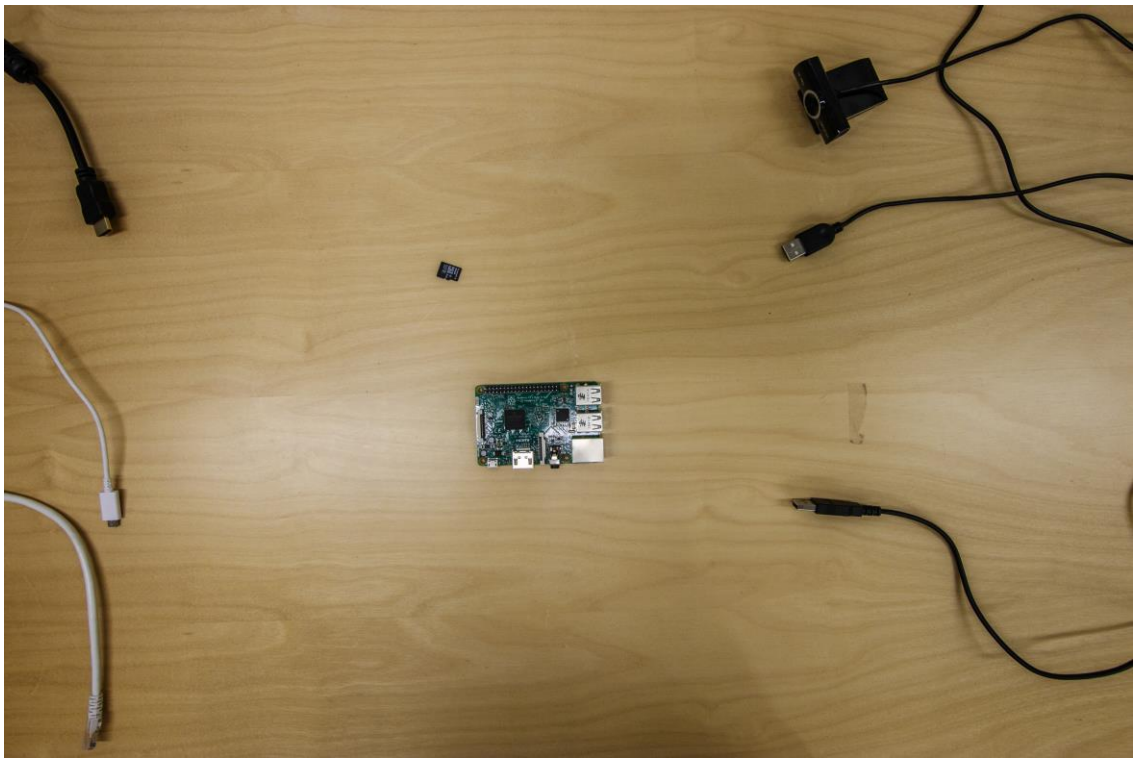

Figure 7. Experimental setting with Rasberry Pi, different cables, and memory card.

A laptop Acer Aspire E5-574G with a screen size 15" and 1366 x 768 resolution was used with a separate mouse. A Tobii X-series X2-60 eye-tracker (Tobii, 2016) combined with Tobii analysis software (Tobii, 2017c) was used for measuring the participants' gaze and analyzing it (see Figure 8).

Figure 8. The setup involved in the study, with a test recording open.

To analyze the scale of error the tracker had for each individual participant TraQuMe (Akkil et al., 2014) was used to calculate the deviance of fixations from each corner of the screen and the center of the screen.

The focus of the study was the evaluation of the user interface of the Pointr application (Pointr, 2017), created by Delta Cygni Labs (see Figure 9). The application Pointr is a remote collaboration tool. It uses video calls and augmented reality as a means of collaboration between two users. The user can call another person and receive help remotely or vice versa the user can call another person to help them. The way the application works is that the two users both see the same screen when in a video call. Thus, the other user can show the helping user what he or she sees, by pointing the camera of his or her mobile device towards the problem they are encountering. The helping user can then use the "pointers" in the application to show what to do via the video feed, in real time (see Figure 10 for an example of pointing). Showing the other user what to do via the video feed can be understood as augmented reality. This then is meant to enable the other user to better understand what to do, in contrast to just having audio feedback as in other forms of video calls.

Figure 9. Pointr application starting screen.

To collaborate with the assistant, a Samsung Galaxy A3 smartphone with a 720 x 1280-pixel screen was used by the assistant to send a live video feed using the smartphone's rear camera and act upon the instructions displayed to him on the screen.

The first two tasks were given to the participants using pieces of paper with tasks written on them (see Appendix B). A background questionnaire was used to gather participant demographics, information on previous experience of similar applications, or the application, as well as, information on whether they have normal vision (see Appendix D). To gain insight into participants' thoughts, while they were looking at the recordings with their gaze pattern overlaid, the laptop's microphone was used to record the think-aloud procedure with the program Audacity version 2.1.2 (Mazzoni, 2017). Additionally, the "user experience questionnaire" (Laugwitz et al., 2008, see Appendix E), which uses a seven-point scale, was used to measure the user experience of the participants.

4.3. Procedure

Upon arrival to the usability lab at the University of Tampere the participants were greeted and asked to have a seat. Participants were introduced to the procedure and asked to sign an informed consent form. After signing the consent form the participants

were taken to the gaze lab, which is next to the usability lab. They were introduced to the assistant, who remained in the gaze lab for the duration of the experiment.

Participants were then shown the Raspberry Pi, and instructed on how to connect each cable and the memory card to the minicomputer. The cables were then removed and the participants were then asked to practice instructing, by pointing out to the researcher where each of the cables and memory card can be attached, by using any way that was natural for them, including speaking and pointing to appropriate sockets on the Raspberry Pi. The participant was made aware that the assistant would act "dumb" in case they tried to tell the assistant for example to "put the USB-cable into the USB-socket". It was explained that the objective of the study is to observe how the participant uses the application to indicate where the cables go. When the participant felt confident that he or she knew how to instruct the assistant, in how to plug each cable into the Raspberry Pi, the study continued.

Next, the participants were taken back to the usability lab and asked to find a comfortable position from their seat. They were then introduced to the calibration procedure of the gaze tracker, where they first needed to find the optimal distance to the screen, with the help of the researcher. After the optimal distance was determined, participants were introduced to the next phase of calibration where they were expected to follow a moving circle with their eyes. The circle visited each corner and the center of the screen, and the spots were used by the software to automatically calibrate the tracker.

To familiarize the participants with the procedure they were then given an introductory task and the gaze tracking recording was started. The task was to go to google, search for "puppies" or "cars", and find a picture the participant likes the most and open it. They were also told to prepare to describe why they chose that picture. After finding a picture they liked the most, they were asked to find a picture they liked the least and open it. Again, preparing to describe why they disliked the picture the most. Afterwards the participant's recording was stopped and the recording was played back with the gaze pattern overlaid. The participant was asked to practice the thinking aloud method and focus on the experience of using google, the reason for making the choices, and explain why they chose the pictures.

After familiarizing the participant with the procedure, the gaze tracking recording was started again and the first actual task was given to the participant printed on a piece of paper (see Appendix B). The first task was to open and register the application by using it. A separate number and email address was supplied to the participant, which the test server of the application recognized. When the participant completed the task the next task was given on another piece of paper. After the second task, at which point the participant connected remotely with the assistant using the application, the assistant proceeded to give verbal tasks in the form of questions e.g. "Can you help me plug these

cables in to the board?" (see Appendix B). The participant used the application to instruct the assistant the way he or she deemed most useful (see Figure 10 for example; note pointer pointing at the Raspberry Pi).

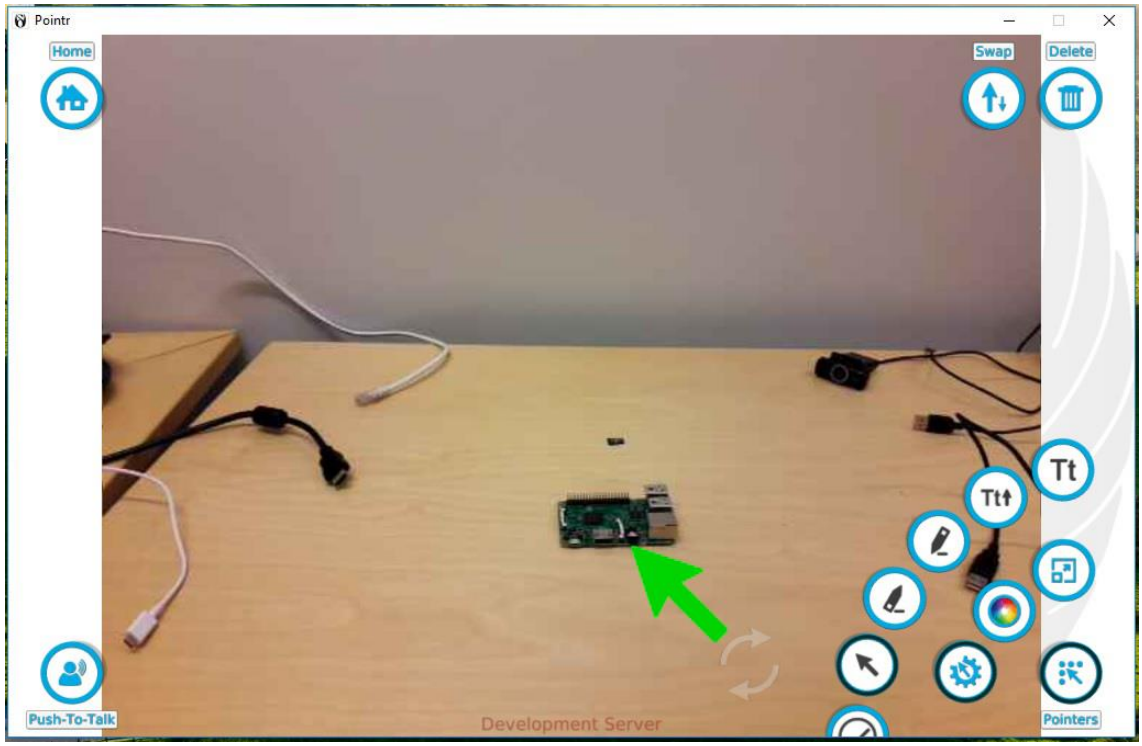

Figure 10. Screen capture from application and the setup. (Not from a real test)

At some point, when about half of the cords were plugged in, the assistant disconnected the call and the researcher verbally asked the participant to re-connect the call, this time without the number. After the participant figured out how to re-connect the call by using contacts, the procedure continued. Finally, the recording was stopped and participants were quidded to use TraQuMe to check the calibration.

Participants were then asked to use the fore learned think-aloud method for the actual test recording. The participants were reminded again to focus on the experience and reasons behind actions taken. The participant's voice was recorded during the playback of the screen recording with the gaze path overlaid. If the participants remained silent for twenty seconds, they were reminded to continue to think aloud by the researcher, e.g. "What are you thinking now?".

Next the participants were asked to fill in the questionnaire about their experience and the background questionnaire. Afterwards the researcher interviewed the participant using a semi-structured interview, meaning that the basic questions were the same but the researcher added additional questions to prompt the participants to give more specific answers (see Appendix B).

4.4. Qualitative analysis method

The objective of the analysis was to find out if the gaze tracking and interview together would result in useful information of the user experience of the application Pointr. The results of these two would be compared to the results of the validated user experience questionnaire. To compare the results the data from the recording would need to be coded into measurable form.

The recordings of each session were analyzed. First, the recordings were transcribed by the researcher. The transcripts contained all the quotes, which were deemed to contain user experience indicating material. In case it was unclear at this point if any specific quote would be useful or if it would alter the meaning of other quotes, the quote was also included in the transcript. The quotes deemed irrelevant, were not included.

After transcriptions were ready for each participant, a plan on how each quote should be coded was made. To be able to compare the results from the recording with that of the questionnaire, the same six descriptive terms were used (Attractiveness, efficiency, perspicuity, dependability, stimulation, and novelty). The rules by which a quote would be associated to any specific term were carefully designed based on the User experience questionnaire's handbook (Schrepp, 2015) and additional rules of conduct in case it was not clear to, which term each quote belonged.

The definition of the attributes was based on User experience questionnaire (Schrepp, 2015) and were as follows:

- Attractiveness: "Overall impression of the product. Do users like or dislike the product?"
- Efficiency: "Can users solve their tasks without unnecessary effort?"
- Perspicuity: "Is it easy to get familiar with the product? Is it easy to learn how to use the product?"
- Dependability: "Does the user feel in control of the interaction?"
- Stimulation: "Is it exciting and motivating to use the product?"
- Novelty: "Is the product innovative and creative? Does the product catch the interest of users?" (Schrepp, p. 2).

Attractiveness is understood as a "pure valence dimension" (Schrepp, p. 2), meaning that it does not have practical or hedonic qualities, which could be distinguished, but it affects both. Efficiency, perspicuity, and dependability related quotes were seen as belonging to pragmatic qualities of the application. Stimulation and novelty related quotes on the other hand were seen as relating to hedonic qualities of the application.

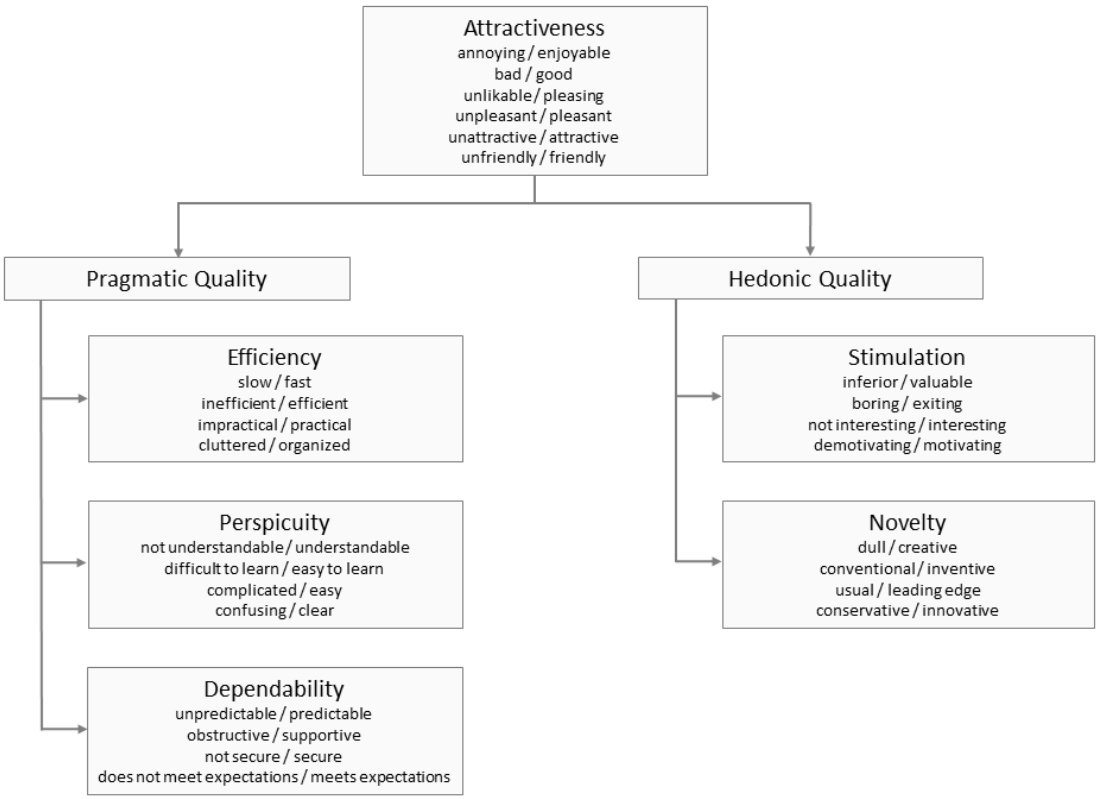

Figure 11. Assumed scale structure of the UEQ. (Schrepp, 2015 p. 3)

The basic way of categorizing a quote to one of the categories was to see if it had a clear indication of belonging to one of the items listed in the questionnaire (see Figure 11 for items under each term). The judgments of clear indications were made by the researcher subjectively, but were based on semantic meanings of the quotes, as well as, matching words. For example, in case the subject said “Finding the button was confusing”, the quote would be associated with perspicuity, because the word “confusing” was used and the sentence can be understood so that something was confusing. Another form of clear indication was if the participant used words like “liked” or “disliked”, meaning that the quote could be judged as positive or negative. All the quotes could not be directly associated with any of the items in any of the categories. Therefore, additional rules for each of the categories were developed to keep the categorization consistent.

For attractiveness, any mention of what the application looked like, without reference to a practicality or experience would be marked as attractiveness. Quotes indicating effectiveness included the reference to time, like “This takes so long”. Also, if the participant expressed that they knew they should do something, but they were not

doing it, because they forgot or it caused too much trouble, it was marked as negative efficiency. If the opposite was true, that the participant found it very easy to use something, because it was always available, it was marked as positive efficiency.

Any practical issues which indicated that the person did not know or knew what to find and experienced that it was difficult or easy was marked as perspicuity. In the case that participants were expecting something to happen when they did something, but did not clearly say that they were expecting it, the quote would still be marked as dependability based on the evaluation by the researcher. If the participant on the other hand said that they could not find something, indicating that they expect to find something, it was marked as negative dependability. Sometimes users were lost but did not say so directly, therefore when a participant said they could not find something the first mark would go to negative dependability, meaning that the application is not predictable, but if the participant said they cannot find something twice it would be marked as negative perspicuity for complicated or not understandable. Some quotes could have a double meaning and refer to two different experiences, in these cases both categories would get a point.

If the participant would try to describe something by using an emotionally charged reference like “that was like pam!”, the quote would be marked as positive or negative stimulation, depending on what impression the quote gave. Also in the case that something was annoying or distracting the participant, it was marked as negative stimulation, unless it was distracting, because the participant found it interesting. Any positive emotional experiences mentioned also got marked as stimulation.

Mentions of regularity were marked as negative novelty. However, if the participant said that they could see this used somewhere where it is not, or that it would make things easier, it would be marked as positive novelty. Judgments on novelty were not made as easily as other forms of experience. Only clear indications were considered.

Many of the judgments were made based on the indication that the participant referred to an experience tied to an item in the category. In unclear cases, the evaluator used the best of his judgment to place the quote in some category. Rest of the judgements were made based on the rules mentioned before.

5. Results

The results comparing the user experience questionnaire (Laugwitz et al., 2008), with the combination of the retrospective think-aloud procedure cued with gaze tracking and the semi-structured interview, are presented in this chapter.

5.1. Data evaluation

Initially the data recorded with TraQuMe (Akkil et al., 2014), was analyzed for any strong discrepancies (see Table 2). The offset of each participant for the average centimeter offset from measurement points was compared to other participant's values. Only one participant's average was above one standard deviation (*sd* = 0,73 cm) from the average of all the participant's means.

Participant	Point 1 offset (cm)	Point 2 offset (cm)	Point 3 offset (cm)	Point 4 offset (cm)	Point 5 offset (cm)	Participant mean (cm)
1	0,82	0,79	0,62	1,17	1,54	0,99
2	0,98	1,46	0,78	1,05	0,91	1,04
3	0,71	1,25	0,15	0,29	1,06	0,69
4	0,62	0,93	0,36	0,37	0,76	0,61
5	0,88	0,77	0,20	0,36	1,09	0,66
6	0,29	0,90	0,69	0,60	0,66	0,63
7	0,67	0,86	0,40	0,50	0,42	0,57
8	2,21	2,96	3,62	4,37	1,74	2,98
9	1,12	0,98	0,26	0,33	0,40	0,62
10	0,51	0,38	1,18	0,70	1,36	0,83
Point mean (cm)	0,88	1,13	0,83	0,97	0,99	Average for all: 0,96 cm

Table 2. Combined eye offset in centimeters

The data for participant 8 showed a higher offset than the data of other participants. The participant's recording was then inspected to see if the gaze tracking seems off. The recording showed slightly offset fixations, but it could still be predicted all the time at which button or area the participant was looking at. To further analyze the need to reject

the participant's data, a heatmap was created (see Figure 12). The heatmap shows, that most gaze activity has remained inside the application. Therefore, the conclusion was made that the participant's data can be considered in the analysis.

Figure 12. Heatmap showing gaze activity of participant 8

5.2. Data from the retrospective think-aloud and the semi-structured interview

The number of different quotes was arranged depending on the attribute the quote belonged to, if it was positive or negative, and if it was said during the retrospective think-aloud or the semi-structured interview. The means and standard deviations for each different measure was calculated (see Table 3).

		Positive mean (*sd*)	Negative mean (*sd*)
Retrospective think-aloud	Attractiveness	0.2 (*0.4*)	0.5 (*0.7*)
	Efficiency	0.7 (*1.1*)	1.9 (*1.3*)
	Perspicuity	2.0 (*1.7*)	6.5 (*2.1*)
	Dependability	0.6 (*0.7*)	4.1 (*2.6*)
	Stimulation	1.2 (*1.8*)	2.1 (*1.8*)
	Novelty	0.2 (*0.6*)	0.2 (*0.4*)
Semi-structured interview	Attractiveness	1.2 (*1.1*)	1.3 (*1.2*)
	Efficiency	0.7 (*0.8*)	0.5 (*0.7*)
	Perspicuity	1.6 (*1.2*)	4.2 (*1.2*)
	Dependability	0.1 (*0.3*)	2.6 (*2.1*)
	Stimulation	2.2 (*1.9*)	1.0 (*0.8*)
	Novelty	0.6 (*0.5*)	0.5 (*0.8*)

Table 3. Means and rounded standard deviations of the number of quotes for each measure.

To compare the results with those of the user experience questionnaire (Laugwitz et al., 2008), the means and standard deviations of the negative measures were adjusted for their weight by multiplying the number of negative quotes with -1. The positive and negative quote means and standard deviations, were then averaged to form means for the different attributes (see Table 4).

All quotes		
Attribute	Mean	Standard deviation
Attractiveness	-0.10	1.28
Perspicuity	-1.78	4.02
Efficiency	-0.25	1.45
Dependability	-1.50	2.56
Stimulation	0.08	2.32
Novelty	0.03	0.73

Table 4. Rounded means and standard deviation of all the quotes adjusted for their weights

Retrospective think-aloud quotes				**Semi-structured interview**		
Attribute	Mean	Standard deviation		Attribute	Mean	Standard deviation
Attractiveness	-0.15	0.67		Attractiveness	-0.05	1.70
Perspicuity	-2.25	4.73		Perspicuity	-1.30	3.20
Efficiency	-0.60	1.76		Efficiency	0.10	0.97
Dependability	-1.75	3.04		Dependability	-1.25	2.02
Stimulation	-0.45	2.42		Stimulation	0.60	2.16
Novelty	0.00	0.56		Novelty	0.05	0.88

Table 5. Retrospective think-aloud and semi-structured interview quote means and standard deviations for different attributes.

To determine if either the means and standard deviations of the retrospective think-aloud method or those of the semi-structured interview affected the results more, their weight-adjusted means and standard deviations were calculated independent of the other results (see Table 5).

5.3. Results from the user experience questionnaire

The answers of each participant from the user experience questionnaire were counted and inserted into the readymade excel calculator included with the user experience questionnaire (Laugwitz et al., 2008). The calculator then automatically calculated the results for the questionnaire. The mean results for each attribute was given (see Figure 13). The results show that most of the results reflect a neutral evaluation (values between -0.8 to 0.8), and stimulation reflects a slightly positive evaluation (values over 0.8).

UEQ Scales	
Attractiveness	0,383
Perspicuity	-0,725
Efficiency	0,600
Dependability	0,475
Stimulation	0,950
Novelty	0,650

Figure 13. Means of the user experience questionnaire attributes as given by the user experience questionnaire calculator, included with the questionnaire.

5.4. Comparing the results

To compare the results the means of all the weight adjusted quotes and the means of the user experience questionnaire were tested for correlation. The correlation coefficient was $r \approx 0.756$ ($r^2 \approx 0.572$). To understand the relation, a line chart of the two was made (see Figure 14). It can be derived from the line chart that each of the means of the quotes are lower in comparison to the means from the user experience questionnaire. The relation between the two is also visible from the chart.

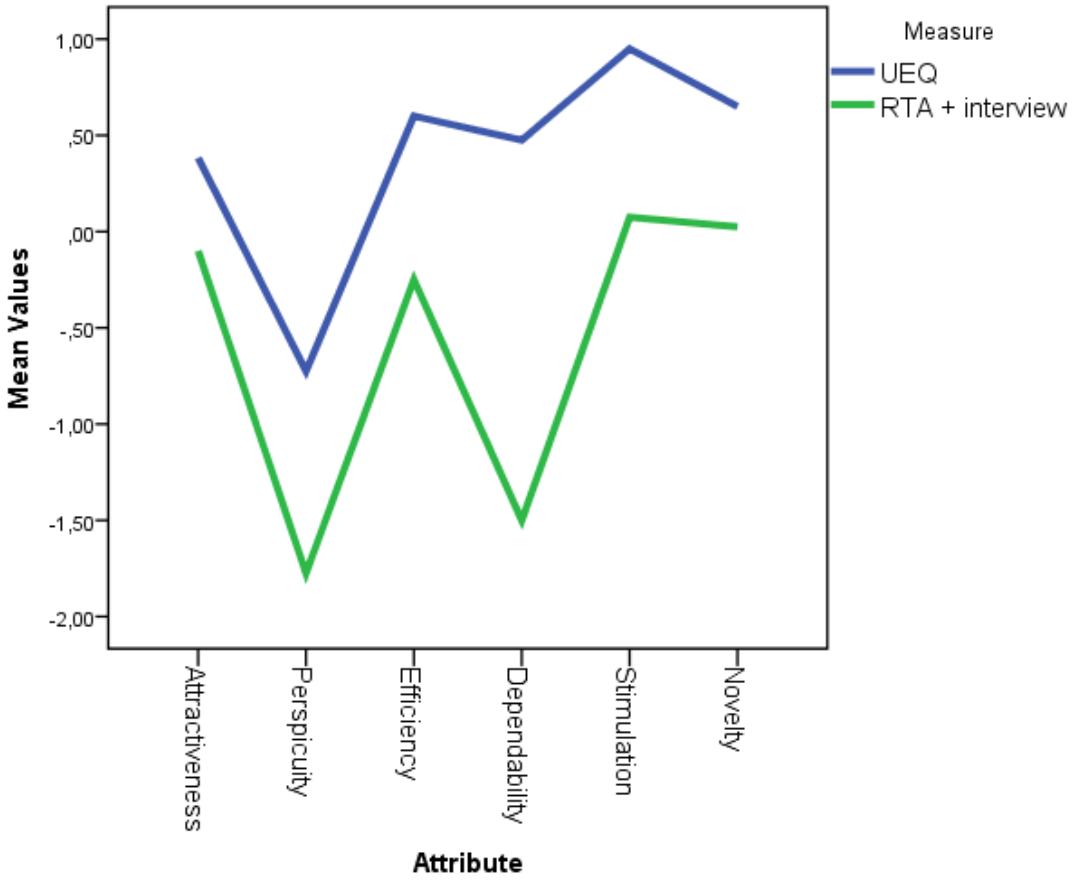

Figure 14. Weight-adjusted means of each attribute for all the quotes plotted on a line chart.

When compared independently, the quotes from the retrospective think-aloud correlate with the user experience questionnaire by $r \approx 0.745$ ($r^2 \approx 0.555$). The main difference can be seen from the line chart, where the means for perspicuity, efficiency, dependability, and stimulation are lower when the quotes from the semi-structured interview are not taken into consideration (see Figure 15).

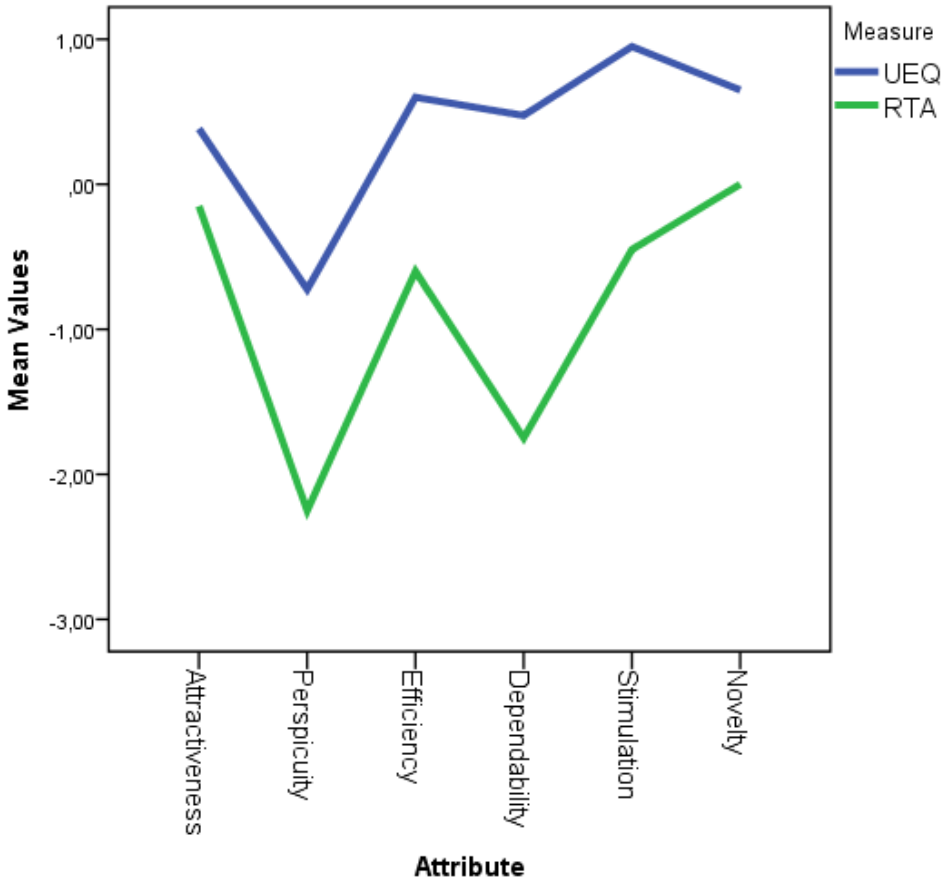

Figure 15. Weight-adjusted means of the quotes from the retrospective think-aloud for each attribute plotted on a line chart.

When compared independently, the quotes from the semi-structured interview correlate with the user experience questionnaire by $r \approx 0.756$ ($r^2 \approx 0{,}572$). The main difference can be seen from the line chart, where opposite to the analysis, retrospective think-aloud is considered independently, the means for perspicuity, efficiency, dependability, and stimulation are higher (see Figure 16).

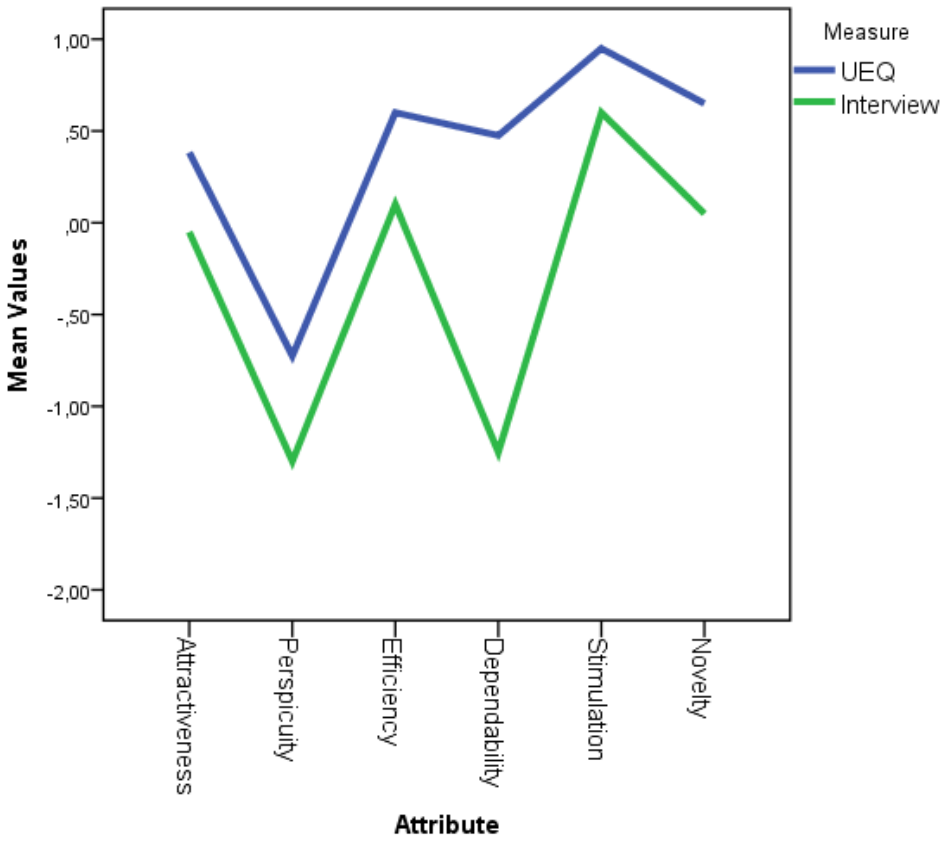

Figure 16. Weight-adjusted means of the quotes from the semi-structured interview for each attribute plotted on a line chart.

The study also produced detailed findings of the potential problems involved with the user interface of the application Pointr. These findings included usability problems as well as user experience issues. These findings were reported directly to the company and are not part of the thesis, because the focus of the thesis is the method itself and its contribution to the evaluation of user experience.

6. Discussion

The objective of the thesis was to assess the possibility of using gaze tracking for the evaluation of user experience. The chosen method required additional means of collecting data, therefore the retrospective think-aloud method and a semi-structured interview were used to collect subjective data from the participants and to evaluate their experience of using the application Pointr (Pointr, 2017). Additionally, a validated user experience questionnaire (Laugwitz et al., 2008) was used as a means of comparison for the data that was acquired. The results of the study confirm that user experience can be measured using gaze tracking combined with the retrospective think-aloud method and an interview. Data acquired correlates with that of the validated user experience questionnaire and adds to what it cannot explain alone.

The correlation between the validated user experience questionnaire and both the retrospective think-aloud combined with gaze tracking and semi-structured interview is surprisingly high. This is probably due to the same concepts being used as the base for the analysis, and it suggests that the measures worked as they were expected to. When combining the results from the two methods, the amount of explained variance is higher than when comparing the methods separately with the user experience questionnaire, as expected. The qualitative analysis, however, tells more than the quantitative indication that the two are related. From the qualitative analysis, insight into the participants' user experience can be derived. The quotes recorded while using the methods can be analyzed and give a detailed view on what the participants experience.

The results of the current study indicate that the methods applied give a more negative picture of the user experience, than the user experience questionnaire, for the application Pointr. However, the results must be analyzed with careful consideration. It is presumed that the method gives a more extreme indication of user experience, which can be more useful in testing software, than that of general user experience measures similar to the user experience questionnaire.

Even when only comparing the validated user experience questionnaire with the retrospective think-aloud, the correlation remains high. This is also true when comparing the questionnaire alone with the semi-structured interview. Moreover, it is important to notice that there is a difference in the qualitative results of the different forms of measures. As Law (2011) described, the argument between qualitative and quantitative measures is an ever-on-going argument, but when accepting the relevance and possibilities involved in both, the outcome is that several types of measures, be they quantitative or qualitative, might complement each other, as they do in this case. Therefore, using any one of the methods alone might not be sufficient in analyzing the user experience of an application.

The results of the retrospective think-aloud method combined with gaze tracking indicate that the most negative quotes on user experience and mostly those of practical quality can be found with this method. However, it is unclear from whether the application elicited such a negative user experience, or if the method itself tends to favor negative quotes. Judging from the available data, it makes sense to presume that the retrospective think-aloud method gives more precise and therefore extreme quotes on specific user experience qualities of the product. Consequently, if the product would have had many exciting features, the quotes from the retrospective think-aloud would reflect the fact by generating an extremely positive picture of the product. That is why the quotes received from this method should be treated independent: as individual quotes that give good indication of the user experience of specific product features.

In addition, the retrospective think-aloud method combined with gaze tracking produced more quotes on practical user experience than the semi-structured interview. This suggests that the retrospective think-aloud helps the participants to indicate specific problems or likings, which seem to be indicative of practical user experience. The study, however, cannot distinguish if quotes indicating practical user experience are thought of as primary issues and are therefore told during the retrospective think-aloud. Consequently, if the interview would have been first, the primary issues might have been told during the interview.

It is difficult to judge to what extent the gaze pattern overlaid on the recording or the recording itself affected the results of the retrospective think-aloud. The study did not distinguish if retrospective think-aloud alone could have produced the same responses from the participants or possibly even more useful responses. Therefore, it cannot be concluded that gaze tracking offered additional benefit in addition to retrospective think-aloud alone. Nevertheless, it is arguable that the cues created by the gaze pattern could have helped the participants to know what they were doing at each point.

Formerly used in usability evaluations (e.g. Guan et al., 2006; Hyrskykari et al., 2008), the stimulated form of retrospective think-aloud has been deemed useful in most cases. Therefore, gaze tracking and retrospective think-aloud compared with retrospective think-aloud separately was not tested in this pioneering study. It is, however, important to note that some studies have found that the gaze pattern overlaid on the recording can be distracting and confronting to the participants (e.g. Elling et al., 2011). Distractions might occur, because participants are not used to seeing their own gaze pattern on the screen and this might affect their attention (Elling et al., 2011). Participants are typically not aware of all the places they looked at meaning that they could possibly get confused or invent something to make sense of their own gaze.

Nevertheless, in an application, which is as static as the application studied (the movement of the interface itself was limited), the added benefit for the participant of

knowing where they are looking can be reasoned to be more valuable than in situations where the interface is less static. Further research is, however, needed to establish if this assumption is accurate and when does using gaze tracking become more of a distraction than a helpful indicator, assuming it does in user experience evaluations also.

The results of the semi-structured interview, on the other hand, give a broader picture of the user experience qualities related with the product. The semi-structured interview also generated negative quotes on average, similar to the retrospective think-aloud method combined with gaze tracking. The difference being that the average was not as extreme.

Additionally, the semi-structured interview generated more hedonic user experience quotes, which were also more positive, in comparison to the retrospective think-aloud. This suggests that when participants are not occupied with explaining what they were thinking, they think of the wider implications and reflect on their overall experience, which in this case was more positive regarding the hedonic user experience.

Unexpectedly, participants also gave more quotes on the attractiveness of the application in the interview, while they were not actively looking at it. However, this finding can possibly be explained by the fact that a question was specifically asked about the visual impression of the application (see Appendix F). Therefore, the application might have been deemed conventional or usual and therefore might not have generated as many quotes on its attractiveness, when not explicitly asked about it.

It is worth noting that, including the interview in the results to improve the amount of variance the methods combined explain is not necessary. It is, however, clear that the qualitative benefit of the interview is beneficial in understanding the factors affecting the overall experience of the product. The quotes from the interview explain what affect the participants' experience. The participant can give their own subjective view on *why* they believe the product was experienced as it was. This highlights the key difference with the user experience questionnaire and the retrospective think-aloud method.

The presumed extent to whether the quotes are more negative or positive are tied to the valence of different extreme experience eliciting features. Judging the overall user experience from the interview might again give a biased view of the overall product experience and the quotes would therefore better be treated as indicators to what cause the overall experience or affect the overall experience rather than the overall experience itself.

The broader objective of the study was to develop a means of studying user experience separated from usability, when using gaze tracking. Prior research on the subject does not exist or is scarce, to the best of my knowledge, and the need for a method that separates usability from user experience when using gaze tracking was needed. The current results suggest that this is possible when combining retrospective

think-aloud, considering the differentiation between the two, and using a separation between practical and hedonic user experience as suggested by Hassenzahl (2007).

The study also separated attractiveness as a third class of user experience, which is associated with both practical and hedonic user experience, but is not included in them. However, this was done mainly to make analysis between the user experience questionnaire (Laugwitz et al., 2008) and the two methods, retrospective think-aloud quotes and semi-structured interview quotes, better comparable. Future studies should consider if attractiveness could be included in practical or hedonic user experience and if not, then consider if there is a need for a third division of user experience.

Regardless of the issue of separating the effect of gaze tracking from the setup, the results offer a convincing suggestion that the method of combining gaze tracking and retrospective think aloud can be used in evaluating user experience and is not limited to usability evaluations. It can therefore be concluded that in certain cases the combination can offer unique insight into the user experience of participants, when compared to the user experience questionnaire or the semi-structured interview. The actual benefit depends on the context of use and the needs of the evaluation.

Furthermore, user experience is often understood as a total experience of using a product, but measuring user experience with different methods and at different times can give different results. Therefore, when planning a user experience evaluation with one of the measures used in the study, considering the impact of time is important. The current study cannot answer if the same answers would be given if the order of measuring were switched. However, it suggests that the retrospective think-aloud combined with gaze tracking produces more specific quotes on the user experience, when compared to the user experience questionnaire, and might therefore be subject to change if the measurement time is changed. The results also cannot answer how this might change the evaluation of user experience.

Despite the convincing results, the retrospective think-aloud combined with gaze tracking and the semi-structured interview cannot be taken as an overall indication of the user experience of the application. The participants are particularly advised to voice any concerns or thoughts on the application and therefore the quotes might only mean that those elements cause, in this case, negative user experience and the overall user experience is different. In this specific study, the quotes that the methods generated tended to be more negative than the overall experience measured by the user experience questionnaire. It is also advisable to consider that the results of the user experience questionnaire might have been influenced by the negativity of the retrospective think-aloud, because it was filled in right after.

Other limitations involved in the study were the small number of participants, which caused high standard deviations, and the convenience sample that might not present the main user base of the product. This should be taken into consideration when analyzing

the results, because it could potentially mean that the results are different in other cases or that individual differences are generally high.

The classification of the quotes is also subject to possible subjective bias, because they were classified based on careful qualitative analysis but by only one researcher. This could possibly mean that slight variance in classification could occur if the quotes were revised by another researcher. However, the variance was assumed small enough for the purposes of this study.

Furthermore, when considering whether to use the method introduced in this thesis it is important to consider the time it takes to complete the testing and compare it with the added benefit it produces. Using the validated user experience questionnaire can offer fast results of the overall user experience of the product when the need of the evaluation is to get a quick overview of how users experience the product. This can be especially useful in a preliminary prototyping phase of product development; however, it cannot answer important questions about specific features of the product. Therefore, often when products are far enough in their development, even when users think the product is generally good, there might be issues that do not get noticed using a questionnaire. Interviewing the participant might help in these situations, they might give more detailed answers, which in return leads to better identification of problems, but the answers might still be unspecific. In situations, similar to this the added benefit of the method introduced can be regarded as worth the additional time.

Using gaze tracking and retrospective think-aloud can offer very detailed information on the features of the user interface that affect the user experience. The advantage of the method, compared to other methods, is that it offers unique accuracy to both the user thinking aloud and the researcher analyzing how specific features affect the experience. On the other hand, if the product is going through a fast development phase and features are changed weekly or faster, using the method introduced can be a waste of resources. Therefore, the method introduced can be very useful in situations when the product is almost complete, or when users are not responding to the product as well as expected and pin pointing the problems are hard.

The added benefit of analyzing the user experience of the product helps the researcher to understand why users like or dislike something. This in return can help in correcting problems that cause negative user experiences. The method introduced does not offer straight forward ways of correcting the users' experiences, but it helps in understanding what the users want or need and in so doing helps in correcting problems that occur in the product. Using gaze tracking in combination with the subjective user insight can potentially offer a powerful new way of correcting user experience of diverse products, systems, and services.

7. Conclusion and future considerations

This thesis introduced a method for combining gaze tracking with retrospective think-aloud to effectively evaluate the user experience of applications. The added benefit of the method is in detailed and specific indications of which features or qualities of products cause specific experiences for participants. In addition, the added benefit of semi-structured interviews was discussed and the results were all compared with those of the already validated method of using the user experience questionnaire (Laugwitz et al., 2008).

Two research questions were introduced in the introduction of the thesis and the following section will reflect on them in relation to the results.

1. Can gaze tracking be used to aid in the measurement of user experience of digital products?

The results offer an indication that gaze tracking can be used to aid in measuring user experience of digital products in combination with retrospective think-aloud. However, gaze tracking alone does not refer to the subjective experience of the participant, which is understood as user experience. Therefore, based on the results, there is no evidence to support the possibility that gaze tracking alone could measure user experience.

Furthermore, the study did not separate the effect of gaze tracking on retrospective think-aloud. Thus, gaze tracking and retrospective think-aloud combined might not differ from using retrospective think-aloud separately. In the study, retrospective think-aloud was used as a method to gather user insight and gaze tracking aided the participants in expressing themselves, by reminding or even showing them, where their visual attention was focused at any given time. This was expected to lead to better results as found previously in usability evaluations combining gaze tracking and retrospective think-aloud (e.g. Guan et al., 2006; Hyrskykari et al., 2008).

2. Are there benefits using methods that combine gaze tracking and user experience in comparison to other forms of user experience measures?

The findings of the study suggest that the method of combining retrospective think-aloud with gaze tracking can generate quotes that specify what causes the experience and can therefore help with making decisions based on the results. The study compared the results of the validated user experience questionnaire (Laugwitz et al., 2008), the retrospective think-aloud combined with gaze tracking

and the semi-structured interview, concluding that each measures different aspects of user experience.

The user experience questionnaire gave a broad idea of the overall user experience of the product, and the retrospective think-aloud and semi-structured interview gave more specific ideas of what caused certain experiences. The results suggest that combining the retrospective think-aloud method with a semi-structured interview produces the best results, when evaluating different features of digital application, in comparison to using a user experience questionnaire alone.

There is a very limited amount of previous research on user experience combined with gaze tracking, and this thesis can only represent a starting point for the evaluation of user experience with the aid of gaze tracking. The study suggests that retrospective think-aloud is a viable tool to accompany gaze tracking, but further research for the evaluation of different kinds of applications should be done to verify that the results are not limited to Pointr (Pointr, 2017) or similar applications.

The study presented in the thesis did not test how retrospective think-aloud combined with gaze tracking compares to retrospective think-aloud per se. Therefore, testing this would establish if the results found in usability evaluations combining retrospective think-aloud and gaze tracking (e.g. Hyrskykari et al., 2008) should be expected in user experience evaluations when combining the two. Establishing this would answer if there is any added benefit of using gaze tracking with retrospective think-aloud when evaluating the user experience of a product.

Furthermore, the study presented only examined the possibility of using gaze tracking as an aid to retrospective think-aloud in user experience evaluation, but future studies should examine the possibility of combining other user experience methods with gaze tracking. Likewise, the possibility of combining the data gathered by the gaze tracker in combination with other methods, could yield potentially interesting results. Even using other forms of gaze tracking, such as head worn gaze trackers to evaluate user experience could be considered depending on the product, system, or service analyzed.

The aforementioned suggestions are only a brief inspection of the possibilities involved with gaze tracking. The more affordable and usable the tracking hardware becomes, the more opportunities they present for the evaluation of user experience. Additionally, the better the evaluation methods become, the better user experience of products, systems, and services can be expected.

References

Abran, A., Khelifi, A., Suryn, W., & Seffah, A. (2003). Usability meanings and interpretations in ISO standards. *Software Quality Journal*, *11*(4), 325-338.

All about UX (2017). All about UX: User experience methods. Retrieved January 26, 2017, from *http://www.allaboutux.org/all-methods*

Akkil, D., Isokoski, P., Kangas, J., Rantala, J., & Raisamo, R. (2014). TraQuMe: a tool for measuring the gaze tracking quality. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 327-330). ACM.

Bargas-Avila, J. A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2689-2698). ACM.

Bevan, N. (2009). What is the difference between the purpose of usability and user experience evaluation methods. In *Proceedings of the Workshop UXEM* '09.

Bojko, A. (2005). Eye tracking in user experience testing: How to make the most of it. In *Proceedings of the UPA 2005 Conference*.

Bowers, V. A., & Snyder, H. L. (1990). Concurrent versus retrospective verbal protocol for comparing window usability. In *Proceedings of the Human Factors Society Annual Meeting* (Vol. 34, No. 17, pp. 1270-1274). Sage CA: Los Angeles, CA: SAGE Publications.

Bruun, A., & Ahm, S. (2015). Mind the gap! comparing retrospective and concurrent ratings of emotion in user experience evaluation. In *Human-Computer Interaction* (pp. 237-254). Springer International Publishing.

Chennamma, H. R., & Yuan, X. (2013). A survey on eye-gaze tracking techniques. *arXiv preprint arXiv:1312.6410*.

Chowdhury, G. G., & Chowdhury, S. (2011). *Information Users and Usability in the Digital Age*. London: Facet Publishing.

Cooke, L. (2005). Eye tracking: How it works and how it relates to usability. *Technical Communication*, *52*(4), 456-463.

Delabarre, E. B. (1898). A method of recording eye movements. *American Journal of Psychology*, *9* (4), 572– 574.

Djamasbi, S., Siegel, M., Skorinko, J., & Tullis, T. (2011). Online viewing and aesthetic preferences of generation y and the baby boom generation: Testing user web site experience through eye tracking. *International Journal of Electronic Commerce*, *15*(4), 121-158.

Djamasbi, S. (2014). Eye tracking and web experience. *AIS Transactions on Human-Computer Interaction*, *6*(2), 37-54. Retrieved from: http://aisel.aisnet.org/thci/vol6/iss2/2

Dodge, R., & Cline, T. S. (1901). The angle velocity of eye movements. *Psychological Review, 8* (2), 145– 157.

Dourish, P., & Bell, G. (2011). Introduction. In *Divining a Digital Future : Mess and Mythology in Ubiquitous Computing*.(pp. 1-6) Cambridge, MA, USA: MIT Press. Retrieved from http://www.ebrary.com.helios.uta.fi

Drewes, H. (2010). *Eye gaze tracking for human computer interaction* (Doctoral dissertation). Ludwig Maximilian University of Munich, Munich, Germany.

Duchowski, A. T. (2007). Eye tracking methodology: theory and practice (2nd ed.). London: Springer.

Eger, N., Ball, L. J., Stevens, R., & Dodd, J. (2007). Cueing retrospective verbal reports in usability testing through eye-movement replay. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it* (Vol.1, pp. 129-137). British Computer Society.

Ehmke, C., & Wilson, S. (2007). Identifying web usability problems from eye-tracking data. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it* (Vol. 1, pp. 119-128). British Computer Society. Retrieved from: http://www.bcs.org/upload/pdf/ewic_hc07_lppaper12.pdf

Elbabour, F., Alhadreti, O., & Mayhew, P. (2017). Eye Tracking in Retrospective Think-Aloud Usability Testing: Is There Added Value?. Journal of Usability Studies, 12(3).

Elling, S., Lentz, L., & de Jong, M. (2011). Retrospective think-aloud method: Using eye movements as an extra cue for participants' verbalizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1161-1170). ACM.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, Mass.: MIT Press.

Fitts, P. M., Jones, R. E., & Milton, J. L. (1950). Eye movements of aircraft pilots during instrument-landing approaches. *Aeronautical Engineering Review, 9* (2), 24– 29.

Forrester, J. V., Dick, A. D., McMenamin, P. G., Roberts, F., & Pearlman, E. (2015). The eye: basic sciences in practice. *Elsevier Health Sciences*.

Goldberg, J. H., & Wichansky, A. M. (2003). Eye Tracking in Usability Evaluation: A Practitioner's Guide. In Radach, R., Hyona, J., & Deubel, H. (Eds.). *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 493-516). Elsevier.

Guan, Z., Lee, S., Cuddihy, E., & Ramey, J. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of*

the SIGCHI conference on Human Factors in computing systems (pp. 1253-1262). ACM.

Harrison, C. M. (2008). *Exploring emotional web experience: More than just usability and good design* (Doctoral dissertation, University of York).

Hassenzahl, M. (2007). The hedonic/pragmatic model of user experience. *Towards a UX Manifesto* (pp. *10-14).*

Hassenzahl, M. (2008). User experience (UX): towards an experiential perspective on product quality. In *Proceedings of the 20th Conference on l'Interaction Homme-Machine* (pp. 11-15). ACM.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, *57*(6), 787-795.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.

Horsley, M., Eliot, M., Knight, B. A., & Reilly, R. (2014). *Current trends in eye tracking research*. Cham: Springer.

Hyrskykari, A., Ovaska, S., Majaranta, P., Räihä, K. J., & Lehtinen, M. (2008). Gaze path stimulation in retrospective think-aloud. *Journal of Eye Movement Research*, *2*(4).

Ishimaru, S., Kunze, K., Uema, Y., Kise, K., Inami, M., & Tanaka, K. (2014). Smarter eyewear: using commercial EOG glasses for activity recognition. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (pp. 239-242). ACM.

International Organization for Standardization. (2010a). *Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts* (ISO/DIS 9241-11.2, 3.1). Retrieved from: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:dis:ed-2:v2:en

International Organization for Standardization. (2010b). *Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems* (ISO 9241-210:2010, 2.15). Retrieved from: https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive psychology*, *8*(4), 441-480.

Keskinen, T. (2015). *Evaluating the user experience of interactive systems in challenging circumstances* (Doctoral dissertation). University of Tampere, Tampere, Finland.

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group* (pp. 63-76). Springer Berlin Heidelberg.

Law, E. L. C. (2011). The measurability and predictability of user experience. In *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems* (pp. 1-10). ACM.

Majaranta, P., & Bulling, A. (2014). Eye tracking and eye-based human–computer interaction. In Advances in physiological computing (pp. 39-65). Springer London.

Manabe, H., Fukumoto, M., & Yagi, T. (2015). Direct gaze estimation based on nonlinearity of EOG. *IEEE Transactions on Biomedical Engineering*, *62*(6), 1553-1562.

Mazzoni, D. (2017). Audacity (Version 2.1.2) [Computer software]. Retrieved from www.audacityteam.org

Mirnig, A. G., Meschtscherjakov, A., Wurhofer, D., Meneweger, T., & Tscheligi, M. (2015). A Formal Analysis of the ISO 9241-210 Definition of User Experience. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 437-450). ACM.

Moczarny, I. M., De Villiers, M. R., & Van Biljon, J. A. (2012). How can usability contribute to user experience?: a study in the domain of e-commerce. In *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference* (pp. 216-225). ACM.

Morimoto, C. H., & Mimica, M. R. (2005). Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, *98*(1), 4-24.

Nielsen, J., Clemmensen, T., & Yssing, C. (2002). Getting access to what goes on in people's heads?: reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on Human-computer interaction* (pp. 101-110). ACM.

Nielsen, J. (2012). Usability 101: Introduction to Usability. Retrieved January 05, 2017, from *https://www.nngroup.com/articles/usability-101-introduction-to-usability/*

Norman, D. A. (2013). *The design of everyday things* (Revised and expanded edition). New York: Basic Books.

Nyström, M., Andersson, R., Holmqvist, K., van de Weijer, J. (2013). The influence of calibration method and eye physiology on eye tracking data quality. *Behavior Research Methods*, *45*(1), 272-288.

Olsson, T., Lagerstam, E., Kärkkäinen, T., & Väänänen-Vainio-Mattila, K. (2013). Expected user experience of mobile augmented reality services: a user study in the context of shopping centres. *Personal and ubiquitous computing*, *17*(2), 287-304.

Petrie, H., & Precious, J. (2010). Measuring User Experience of websites: Think aloud protocols and an emotion word prompt list. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (pp. 3673-3678). ACM.

Pointr Easy Remote Support. (2007). Retrieved February 16, 2017, from http://www.deltacygnilabs.com/#products

Poole, A., & Ball, L. J. (2006). Eye tracking in HCI and usability research. In C. Ghaoui (Ed.), *Encyclopaedia of human-computer interaction* (pp. 211-219). Pennsylvania: Idea Group Inc.. Retrieved from: http://www.alexpoole.info/blog/wp-content/uploads/2010/02/PooleBall-EyeTracking.pdf

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of experimental psychology*, *109*(2), 160-174.

Pretorius, M. C., Calitz, A. P., & van Greunen, D. (2005). The added value of eye tracking in the usability evaluation of a network management tool. In *Proceedings of the 2005 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries* (pp. 1-10). South African Institute for Computer Scientists and Information Technologists.

Rajeshkumar, S., Omar, R., & Mahmud, M. (2013). Taxonomies of User Experience (UX) evaluation methods. In *Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on* (pp. 533-538). IEEE.

Ritter, F. E. (2013). *Running behavioral studies with human participants: A practical guide* (1st ed.). Thousand Oaks, CA: Sage.

Romano Bergstrom, J., & Schall, A. (2014). *Eye tracking in user experience design*. Burlington: Elsevier Science.

Roto, V., Law, E., Vermeeren, A., & Hoonhout, J. (2011). User Experience White Paper. *Outcome of the Dagstuhl Seminar on Demarcating User Experience*, Germany. Retrieved from: http://www.allaboutux.org/uxwhitepaper

Roto, V., Obrist, M., & Väänänen-Vainio-Mattila, K. (2009). User experience evaluation methods in academic and industrial contexts. In *Proceedings of the Workshop UXEM'09*.

Russell, J. A. (1978). Evidence of Convergent Validity on the Dimensions of Affect. *Journal of Personality and Social Psychology*, *36*(10), 1152-1168.

Rusu, C., Rusu, V., Roncagliolo, S., Apablaza, J., & Rusu, V. Z. (2015). User experience evaluations: challenges for newcomers. In *International Conference of Design, User Experience, and Usability* (pp. 237-246). Springer International Publishing.

Schrepp, M. (2015). *User Experience Questionnaire Handbook* (2nd ed., pp. 1-12). Retrieved from http://www.ueq-online.org

Tobii AB. (2016). Pro X2-60 screen-based eye tracker. Retrieved December 02, 2016, from http://www.tobiipro.com/product-listing/tobii-pro-x2-60/

Tobii AB. (2017a). Tobii Pro X2 series hero shot. Retrieved January 17, 2017, from http://www.tobii.com/group/news-media/image-gallery/

Tobii AB. (2017b). A woman wearing Tobii Pro Glasses 2 looking at a shelf in the store. Retrieved January 17, 2017, from http://www.tobii.com/group/news-media/image-gallery/

Tobii AB. (2017c). Pro studio software. Retrieved April 28, 2017, from https://www.tobiipro.com/product-listing/tobii-pro-studio/

Tractinsky, N. (2004). Toward the study of aesthetics in information technology. In *Proceedings of the Twenty-Fifth International Conference on Information Systems* (pp. 771–780). AIS Electronic Library

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97-136.

Tuch, A. N., Roth, S. P., HornbæK, K., Opwis, K., & Bargas-Avila, J. A. (2012). Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. *Computers in Human Behavior*, *28*(5), 1596-1607.

Van Den Haak, M., De Jong, M., & Jan Schellens, P. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & information technology*, *22*(5), 339-351.

Vermeeren, A. P., Law, E. L. C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (pp. 521-530). ACM.

Appendix A

CONSENT TO PARTICIPATE IN AND BE RECORDED IN A SOFTWARE TEST

You have been invited to participate in an evaluation of a software that is part of my master's thesis work at the University of Tampere. By participating in the test you will help to evaluate the user experience of the augmented reality application Pointr, developed by Delta Cygni Labs.

You will be asked to perform different tasks using the service while your eyes are tracked. Afterwards, the on screen recording with your gaze pattern overlaid will be played back to you and you will be asked to elaborate on your thought process, by using the think aloud method. In addition, you will be asked to fill in a questionnaire and you will be interviewed about the use of the application.

While you test the software two separate recordings will be made. First a camera will record your eyes and create a gaze pattern which can be analyzed later. Second, the things happening on the screen will be recorded as a video. Meaning that the actions and tasks that take place within the software will also be recorded (i.e. Where and what you click to perform the tests on the software). In the next phase of the study, an additional recording of your voice will be added to enable the analysis of what you said.

The results of the test will be reported anonymously. A summary of the main results will be delivered to the developers of the service. Recordings and participants' personal data will be kept confidential and secured.

You may stop participating in the user experience test at any point with no penalty.

Feel free to ask any questions you may have about your participation.

By signing this form, you will accept the above terms.

Date and place: _____

Signature: _____

Name clarification: _____

Appendix B

Tasks and assistant's script

Tasks given by researcher on paper:

Task 1

- Please start the application and register. Use the number provided and the email address provided to register.

Task 2

-Please establish a connection with the assistant using +358XXXXXXXX (a number was given to participants, which would connect to the assistant)

Assistant's script and tasks

1. **Turn raspberry pi wrong way around on the table.**
2. **Arrange different cables around the table.**
3. **Establishing connection**
    1. "Hello, can you hear me?"
    2. "Can you change the video so that you see my camera?"
4. **Introducing problem**
    1. "I need to connect these cables to this, can you help me?"
    2. "Great! What should I do first?"
5. **Plugging in cables**
    1. Follow instructions, until half of the cables are connected.
    2. Disconnect from the call.
6. **Interruption**
    1. Wait until the participant calls again.
7. **Plugin more cables**
    1. "Hello again, something seems to have gone wrong with the connection, please continue"
    2. Plugin rest of the cables.
8. **Finish**
    1. "Thank you, you can now end the call"

If participant…

- **Uses words like USB, HDMI, or other words that indicate to something:**
    - "Can you show me which one?" or other relevant replys.
- **Leaves the pointers on the screen when they are not needed anymore and starts explaining something else:**
    - "Can you remove the pointers from the screen so it is more clear?"

- **Takes 2-3 minutes with any cable or turning on the raspberry pi and you can see that he or she does not know.**
    - “Do you know what to do next?”
    - If yes, then wait. If no then give some hint, but first try to be sure that they are stuck.

Appendix C

SCRIPT

1. Introduction

- I will first tell you about the study and then ask you to fill in an informed consent, to verify that you understand the point of the study.
- We are conducting a study on analyzing the experience of using the application Pointr by Delta Cygni Labs.
- Pointr is a remote collaboration application.
- Some of the tasks might be difficult, but that is because of the system, not you. Your input is very important to us and there are no wrong answers.
- You will be asked to remotely collaborate with my colleague whom you will meet soon.
- We will be using an eye tracking system. And your eye movements will be recorded and they will be made into visualized data.
- You will also be asked to go through the onscreen recording of your eye movements and elaborate on them.
- Afterwards you will be asked to fill in a questionnaire and some background information.
- At the end, I will ask you some questions about the application.
- You can stop your participation at any point of the study. Do you want to move forward?

2. Handing out informed consent form

- First I'd like you to read this consent form and sign it once you understand it. If you have any questions, please ask me.

3. Getting used to raspberry pi

- Most people probably are not familiar with this mini-computer, so I will introduce it to you.
- Have you seen or used this before?
- This is the raspberry pi, it has multiple ports like the USB port here.
- The ports are all similar to what you would find on a laptop.
- Next I will show you how to plug everything into the raspberry pi and then you will get to do it next.
- It is not difficult it you are used to any laptop or PC, but please try to put all the cables and the memory card in the correct sockets and if you have a question please ask me.
- Later you will be asked to help another person to plug these in, do you think you will be able to or would you like to try again to be sure?

- The other person will know how to plug the cables in, but he will be pretending he does not know, your task will be to figure out a way to instruct him.

4. Calibration

- First we need to calibrate the system
- Please find a comfortable position.

5. Check TraQuMe

- Next we need to check the accuracy

6.Test eye tracking with a webpage

- We will first get you familiar with the procedure
- OK so first, which do you like more cars or puppies?
- I will turn on the recording now and I would like you to go to the google and search for “puppies” (or “cars”).
- Find a picture you like the most and open the picture. After opening it, tell me “I like this one the most” and then you can close that picture. After that please find the picture you dislike the most. Please don’t choose the first one you see, before looking at the other possibilities.

7. Go through recording and try think-aloud method

-Please don’t turn around in order to maintain the calibration.

- Next we will go through the recording and I would like you to practice the think aloud method used in this study.
- -I would like you to tell me what you were thinking while you were navigating.
- -We are not interested in things like “I pressed this button”, but instead we are interested in things like “I saw this button and I thought that this should lead me to where I wanted to go.” or “I thought that this is cool”

8. Start actual study

- We are now going to start the actual study
- I will now put on the recording and I would like you to do this task (give paper and phone number) and inform me when you are done.
- I will then give you another task. I will not be helping you with the tasks, if you have trouble, it is the fault of the system and we want to know if the users are having trouble.
- You can ask me questions, but I might not be able to answer you.

9. Take tasks and phone number away after each task is complete.

10. Stop recording and run TraQuMe

- I will now run the analysis program again in order to see if the accuracy changed between the test.

11. Start the think aloud procedure.

- I will now playback the video for you and you will be able to see your eye movements on the screen as before.

- -I would like you to use the think aloud method while watching the video and tell me what you were thinking while you were looking at certain things.
- Remember, we are not interested in what is happening on the screen, but "for example, 'I was not sure which link to choose, but I chose link x because I thought it was the best match with the information I was looking for" (Elling et al. 2011).
- We are also interested in all your opinions, if you dislike something or like something, please say it while thinking aloud.
- If you remain silent for a time, I will remind you to please tell me what you are thinking.
- If the gaze pattern disappears at some point, it might be because you looked away. So please just continue explaining.
- Let's start the video.

12. Give questionnaire and background questionnaire

- Please fill in this questionnaire and background questionnaire afterwards I will ask you some questions about the application.

13. Interview participant

Appendix D

Background questionnaire

Please circle the appropriate choice.

Age group: <20 20-29 30-39 40-49 50-60 >60

Gender: Male Female Other

Do you use eye glasses? Yes No

Other complications with vision? Yes No

If yes, please specify:

_____

_____

Have you used eye tracking before? Yes No

If yes, please specify:

_____

_____

Have you used remote collaboration tools before? Yes No

If yes, please specify:

_____

_____

Have you used Pointr before? Yes No

If yes, please specify:

_____

_____

Appendix E

Please make your evaluation now

For the assessment of the product, please fill out the following questionnaire. The questionnaire consists of pairs of contrasting attributes that may apply to the product. The circles between the attributes represent gradations between the opposites. You can express your agreement with the attributes by ticking the circle that most closely reflects your impression.

Example:

attractive	○	⊗	○	○	○	○	○	unattractive

This response would mean that you rate the application as more attractive than unattractive.

Please decide spontaneously. Don't think too long about your decision to make sure that you convey your original impression.

Sometimes you may not be completely sure about your agreement with a particular attribute or you may find that the attribute does not apply completely to the particular product. Nevertheless, please tick a circle in every line.

It is your personal opinion that counts. Please remember: there is no wrong or right answer!

Please assess the product now by ticking one circle per line.

	1	2	3	4	5	6	7		
annoying	○	○	○	○	○	○	○	enjoyable	1
not understandable	○	○	○	○	○	○	○	understandable	2
creative	○	○	○	○	○	○	○	dull	3
easy to learn	○	○	○	○	○	○	○	difficult to learn	4
valuable	○	○	○	○	○	○	○	inferior	5
boring	○	○	○	○	○	○	○	exciting	6
not interesting	○	○	○	○	○	○	○	interesting	7
unpredictable	○	○	○	○	○	○	○	predictable	8
fast	○	○	○	○	○	○	○	slow	9
inventive	○	○	○	○	○	○	○	conventional	10
obstructive	○	○	○	○	○	○	○	supportive	11
good	○	○	○	○	○	○	○	bad	12
complicated	○	○	○	○	○	○	○	easy	13
unlikable	○	○	○	○	○	○	○	pleasing	14
usual	○	○	○	○	○	○	○	leading edge	15
unpleasant	○	○	○	○	○	○	○	pleasant	16
secure	○	○	○	○	○	○	○	not secure	17
motivating	○	○	○	○	○	○	○	demotivating	18
meets expectations	○	○	○	○	○	○	○	does not meet expectations	19
inefficient	○	○	○	○	○	○	○	efficient	20
clear	○	○	○	○	○	○	○	confusing	21
impractical	○	○	○	○	○	○	○	practical	22
organized	○	○	○	○	○	○	○	cluttered	23
attractive	○	○	○	○	○	○	○	unattractive	24
friendly	○	○	○	○	○	○	○	unfriendly	25
conservative	○	○	○	○	○	○	○	innovative	26

Appendix F

Semi-structured interview questions

1. How was it? (casual question to get started)
2. What is your overall experience of the application?
3. Did you have difficulties in some tasks? (if yes, can you elaborate on them?)
4. What do you think about the registration procedure of the application?
5. How about making a call?
6. How would you describe the experience of giving advice to the other person?
7. What do you think about the visual impression of the application?
8. What was the best thing about the application?
9. What was the worst thing about the application?
10. How would you improve it?
11. Anything you would like to ask?